

Voice Activity Detection Based on Multiple Statistical Models

Joon-Hyuk Chang, *Member, IEEE*, Nam Soo Kim, *Member, IEEE*, and Sanjit K. Mitra, *Life Fellow, IEEE*

Abstract—One of the key issues in practical speech processing is to achieve robust voice activity detection (VAD) against the background noise. Most of the statistical model-based approaches have tried to employ the Gaussian assumption in the discrete Fourier transform (DFT) domain, which, however, deviates from the real observation. In this paper, we propose a class of VAD algorithms based on several statistical models. In addition to the Gaussian model, we also incorporate the complex Laplacian and Gamma probability density functions to our analysis of statistical properties. With a goodness-of-fit tests, we analyze the statistical properties of the DFT spectra of the noisy speech under various noise conditions. Based on the statistical analysis, the likelihood ratio test under the given statistical models is established for the purpose of VAD. Since the statistical characteristics of the speech signal are differently affected by the noise types and levels, to cope with the time-varying environments, our approach is aimed at finding adaptively an appropriate statistical model in an online fashion. The performance of the proposed VAD approaches in both the stationary and nonstationary noise environments is evaluated with the aid of an objective measure.

Index Terms—Discrete cosine transform (DCT), generalized gamma function, maximum likelihood.

I. INTRODUCTION

VOICE activity detection (VAD) refers to the classical problem of distinguishing active speech from nonspeech and has applications for a variety of speech communication systems, such as speech coding, speech recognition, noisy speech enhancement, hands-free conference, and echo cancellation. Particularly, VAD has been an essential part of variable-rate speech coding which is one of the most elegant ways of enhancing the capacity and coverage of communication bandwidth. For that reason, various types of VAD algorithms have been proposed. Earlier algorithms for VAD are mostly based on the linear prediction coding (LPC) parameters [1], energy levels, formant shape [2], zero crossing rate (ZCR) [3], cepstral feature [4], adaptive modeling of voice signals [5], and the periodicity measure [6]. More recently, the VAD approaches based on a pattern recognition [7] and higher order cumulants

of the LPC residual have been presented as new strategies [24]. Specifically, for example, the energy difference, ZCR, and spectral difference are applied to ITU-T G.729 Annex B [9]. A similar approach is applied to the selectable mode vocoder in 3rd Generation Partnership Project2 (3GPP2) [10], ITU-T G. 723. 1 [11], and European Telecommunications Standards Institute Adaptive Multi-Rate (AMR) VAD option 2 [12].

Recently, efforts to enhance the performance of VAD by employing a statistical model have been made with the decision rule being derived from the likelihood ratio test (LRT) applied to a set of hypotheses. The statistical model-based VAD algorithm originated from Ephraim and Malah's work [8]. Sohn *et al.* applied a Gaussian statistical model to VAD [13] in which the estimation of parameters is conducted by means of the decision-directed (DD) approach [14]. This algorithm has shown a high detection accuracy compared with the conventional algorithms in spite of the need for optimization of a few relevant parameters. The statistical model-based method has been further improved with the incorporation of the soft decision scheme [15], [16]. As is well known, most of the conventional VAD algorithms which mainly operate in the discrete Fourier transform (DFT) domain assume that the clean speech and noise spectra are characterized by Gaussian distributions. Recently, it has been reported that the clean speech and noise DFT coefficients are more effectively described in terms of the probability density functions (pdfs) such as the Gamma and Laplacian distributions [31], [32].

One of the key issues in VAD is robust estimation of the signal-to-noise ratio (SNR), and many studies have been performed in the past to this end. In [14], the DD estimation scheme is applied to compute the *a priori* SNR which was originally proposed in [8] to reduce the bias incurred from the maximum likelihood (ML) estimation of the spectral amplitude. The DD method has proven to provide a smoother estimate for the *a priori* SNR than the ML approach during the nonspeech periods. The general behavior of this DD approach was comprehensively analyzed in [30], where it has also been shown that the DD approach reduces the musical noise quite effectively. Another innovation in VAD is the incorporation of soft decision into power spectrum estimation [16]–[18]. Other successful approaches have been developed to provide a robust noise estimate through the minimum noise statistics and minima controlled recursive averaging [21], [33].

In this paper, several candidate models are investigated to characterize the distribution of the spectral components under various noisy conditions. Not only the traditional Gaussian pdf but also the complex Laplacian and Gamma pdfs are applied to

Manuscript received June 10, 2004; revised June 27, 2005. This work was supported in part by IITA through IT Leading R&D Support Project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andrew C. Singer.

J.-H. Chang is with the Department of Electronic Engineering, Inha University, Incheon 402-751, Korea (e-mail: changjh@inha.ac.kr).

N. S. Kim is with the School of Electrical Engineering and the Institute of New Media and Communications, Seoul National University, Seoul 151-742, Korea.

S. K. Mitra is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

Digital Object Identifier 10.1109/TSP.2006.874403

represent the distribution of each DFT coefficient. Each parametric model is evaluated via the goodness-of-fit (GOF) test which measures how far the assumed model deviates from the empirical distribution. Among many possible GOF tests, we employ the *Kolmogorov–Smirnov* (KS) test [28]. Based on the analysis of the KS tests, we have found that the newly proposed Laplacian and Gamma models are more suitable for modeling the noisy speech spectra in most noise conditions. The LRT is derived from each parametric model, in which the decision rule is described in terms of the relevant parameters. This idea was originally proposed in [19], and this paper gives an in depth description as well as extensive experimental results with the Gamma model. Moreover, we propose new approaches which further improve the original work under various noise conditions as described below.

First, we introduce the *smoothed form* for the likelihood ratio (LR) which makes it possible to improve the performance of the LR compared with the one proposed in [19]. With a tradeoff between the instantaneous noisy power and its estimated value, we construct the smoothed form of the LR. The second approach is a predictive (PD) estimation which renders the *a priori* SNR to be more precise and reliable. In order to overcome some shortcomings of the earlier work [8], [14], the employed PD estimation incorporates the soft-decision scheme which enables a robust performance of the voice detection algorithm. Moreover, an adaptive way of selecting an adequate statistical model for the different noise types and SNRs is proposed based on the online KS test. The approach can be implemented as a simple front of VAD for choosing the corresponding statistical model depending on the given speech frames.

The organization of this paper is as follows. A detailed description of the statistical modeling of noisy speech spectra is given in Section II. The LRT for each parametric model is derived in Section III, and a robust way to estimate the SNR is given in Section IV. In Section V, a number of experiments are conducted to evaluate the performance. In Section VI, some concluding remarks are drawn.

II. STATISTICAL MODELS FOR NOISY SPEECH

We assume that a noise signal n is added to a speech signal s , with their sum being denoted by x . Given two hypotheses H_0 and H_1 , which, respectively, indicate speech absence and presence, it is assumed that

$$H_0 : \text{speech absent} : \mathbf{X}(t) = \mathbf{N}(t) \quad (1)$$

$$H_1 : \text{speech present} : \mathbf{X}(t) = \mathbf{N}(t) + \mathbf{S}(t) \quad (2)$$

in which $\mathbf{X}(t) = [X_0(t), X_1(t), \dots, X_{M-1}(t)]^T$, $\mathbf{N} = [N_0(t), N_1(t), \dots, N_{M-1}(t)]^T$, and $\mathbf{S}(t) = [S_0(t), S_1(t), \dots, S_{M-1}(t)]^T$ are the DFT coefficients at frame t of the noisy speech, noise, and clean speech, respectively.

A. Statistical Models

In this section, we consider three different pdfs as the candidate distributions for statistical modeling. The first one is the complex Gaussian pdf which is most widely applied to charac-

terize the DFT coefficients' distribution in speech analysis [8], [13]–[17], [20], [22], [30], [31]. With the Gaussian pdf assumption, the distributions of the noisy spectral components conditioned on both hypotheses are given by

$$p_G(X_k|H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\} \quad (3)$$

$$p_G(X_k|H_1) = \frac{1}{\pi[\lambda_{n,k} + \lambda_{s,k}]} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k} + \lambda_{s,k}}\right\} \quad (4)$$

where $\lambda_{n,k}$ and $\lambda_{s,k}$ denote the variances of N_k and S_k , respectively.

The second distribution is the complex Laplacian pdf. The real and imaginary parts of each DFT coefficient are assumed to be distributed according to a real Laplacian pdf. Let $X_{k(R)}$ and $X_{k(I)}$ denote the real and imaginary parts, respectively, of the DFT coefficient X_k . If both the real and imaginary parts have the same variance, their distributions are given by

$$p_L(X_{k(R)}) = \frac{1}{\sigma_x} \exp\left\{-\frac{2|X_{k(R)}|}{\sigma_x}\right\} \quad (5)$$

$$p_L(X_{k(I)}) = \frac{1}{\sigma_x} \exp\left\{-\frac{2|X_{k(I)}|}{\sigma_x}\right\} \quad (6)$$

where σ_x^2 denotes the variance of X_k . If the real and imaginary parts of X_k are further assumed to be independent [26], [31],¹ the distribution $p(X_k)$ of X_k turns out to be

$$p_L(X_k) = p_L(X_{k(R)}) \cdot p_L(X_{k(I)}) \quad (7)$$

$$= \frac{1}{\sigma_x^2} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sigma_x}\right\}. \quad (8)$$

From (8), the distributions of the DFT coefficients under the respective hypotheses are given by [19]

$$p_L(X_k|H_0) = \frac{1}{\lambda_{n,k}} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{\lambda_{n,k}}}\right\} \quad (9)$$

$$p_L(X_k|H_1) = \frac{1}{\lambda_{n,k} + \lambda_{s,k}} \times \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{\lambda_{n,k} + \lambda_{s,k}}}\right\}. \quad (10)$$

The last statistical model is described in terms of the complex Gamma pdf. Here, the real and imaginary parts of each DFT coefficient are assumed to be distributed according to a real Gamma pdf according to

$$p_M(X_{k(R)}) = \left(\frac{\sqrt{6}}{8\pi\sigma_x|X_{k(R)}|}\right)^{0.5} \times \exp\left\{-\frac{\sqrt{3}|X_{k(R)}|}{\sqrt{2}\sigma_x}\right\} \quad (11)$$

¹To justify the assumption of independence of real and imaginary parts, we evaluated scatter plots of the real and imaginary parts of DFT coefficients, where the procedure is similar to [31]. From these plots, it is concluded that the real and imaginary parts are only weakly dependent and the error is negligible.

$$p_M(X_{k(I)}) = \left(\frac{\sqrt{6}}{8\pi\sigma_x |X_{k(I)}|} \right)^{0.5} \times \exp \left\{ -\frac{\sqrt{3} |X_{k(I)}|}{\sqrt{2}\sigma_x} \right\}. \quad (12)$$

If the real and imaginary parts are further assumed to be independent of each other as in the Laplacian case, the distribution of a DFT coefficient X_k is then given by

$$p_M(X_k) = \frac{\sqrt{6}}{8\pi\sigma_x |X_{k(R)}|^{0.5} |X_{k(I)}|^{0.5}} \times \exp \left\{ -\frac{\sqrt{3} (|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{2}\sigma_x} \right\}. \quad (13)$$

Applying (13) to the two hypotheses H_0 and H_1 , we have the distributions of the DFT coefficients as follows:

$$p_x M(X_k|H_0) = \frac{\sqrt{6}}{8\pi\sqrt{\lambda_{n,k}} |X_{k(R)}|^{0.5} |X_{k(I)}|^{0.5}} \times \exp \left\{ -\frac{\sqrt{3} (|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{2}\sqrt{\lambda_{n,k}}} \right\} \quad (14)$$

$$p_M(X_k|H_1) = \frac{\sqrt{6}}{8\pi\sqrt{\lambda_{n,k} + \lambda_{s,k}} |X_{k(R)}|^{0.5} |X_{k(I)}|^{0.5}} \times \exp \left\{ -\frac{\sqrt{3} (|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{2}\sqrt{\lambda_{n,k} + \lambda_{s,k}}} \right\}. \quad (15)$$

B. Goodness-of-Fit Test

For a successful VAD operation, we must select a model that fits best the given noisy speech spectra. To this end, we carried out a statistical fitting test for the noisy spectral components conditioned on both H_0 and H_1 under various noise conditions. This is distinguished from the approach proposed by Martin [31] in which the minimum mean-square error (MMSE) estimator is obtained based on a set of clean speech distributions. For the purpose of evaluating each parametric pdf, we apply the KS test which is well-known as the GOF test [27]. Due to the incorporation of the KS test, a reliable survey of each statistical assumption is guaranteed. An attractive feature of this test is that the distribution of the KS test statistic itself does not depend on the underlying cumulative distribution function (cdf) being tested. Another advantage is that it is an exact test, while the chi-square GOF test depends on an adequate sample size for the approximations to be valid.

Speech material of 64 s duration was collected from each of four male and four female speakers.² For simulating noisy environments, the white, vehicular, and babble noises³ from the NOISEX-92 database were added to the clean speech signal at various SNRs. The typical magnitude spectra of the employed noise are displayed in Fig. 1 for easy comprehension of noise

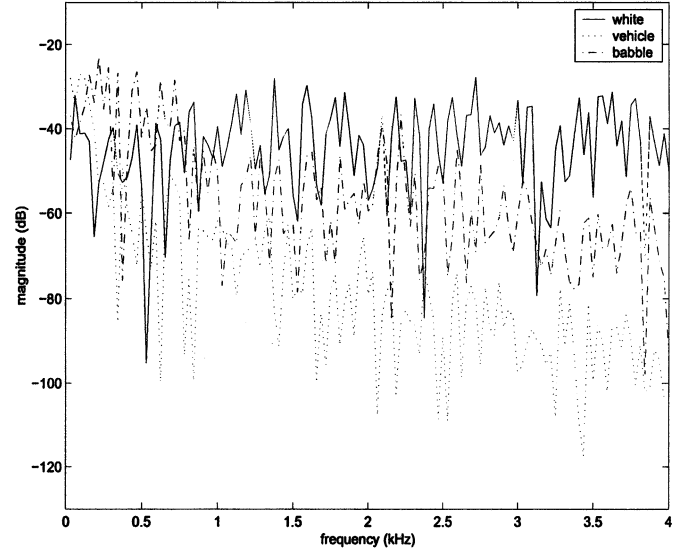


Fig. 1. Typical magnitude spectra of the white, vehicular, and babble noises.

characteristics. A trapezoidal window of length 13 ms was applied to the input signal for every 10 ms. Each sentence was sampled at 8 kHz. Each frame of the windowed signal was then transformed to the corresponding spectrum through a 128-point fast Fourier transform (FFT) after zero-padding. From the collected data, sample mean and variance were computed and then applied to the given parametric distribution.

Let $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$ be a set of the DFT coefficients extracted from the noisy speech data which comprise all frequency bins. The KS test compares the empirical cdf F_X to a given distribution function F . The empirical cdf is defined by

$$F_X(z) = \begin{cases} 0, & z < X_{(1)} \\ \frac{n}{N}, & X_{(n)} \leq z < X_{(n+1)}, \quad n = 1, 2, \dots, N-1 \\ 1, & z \geq X_{(N)} \end{cases} \quad (16)$$

where $X_{(n)}, n = 1, \dots, N$, are the order statistics of the data $\mathbf{X}$ [27]. In order to compute the order statistics, we sort and order the elements of $\mathbf{X}$ so that $X_{(1)}$ represents the smallest element of $\mathbf{X}$ and $X_{(N)}$ is the largest [28].

In order to measure the distance between the empirical cdf and the given distribution, we use the KS statistic [28]. Let $F(\cdot)$ represent a cdf generated by a parametric pdf and $F_X(\cdot)$ be the corresponding empirical cdf of the collected data. Then, the KS test statistic T using the data $\mathbf{X}$ is defined by

$$T(\mathbf{X}) = \max_i |F_X(X_i) - F(X_i)| \quad (17)$$

where the distance is given by the maximum difference between $F_X(\cdot)$ and $F(\cdot)$ evaluated at the sample points $\{X_i\}$. Note that the distribution that yields the smallest KS statistic provides the best fit to the data. An attractive feature of this KS test is that the distribution of the KS test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test, while the chi-square GOF test depends on an adequate sample size for the approximations to be valid.

²A continuous and solid distribution using 64 s speech material for the robust GOF test is obtained (for larger size, results are similar).

³From the time-frequency analysis, it was found that the spectral characteristic of the vehicular noise was quasi-stationary while the spectral shape of the babble noise was rather rapidly changing.

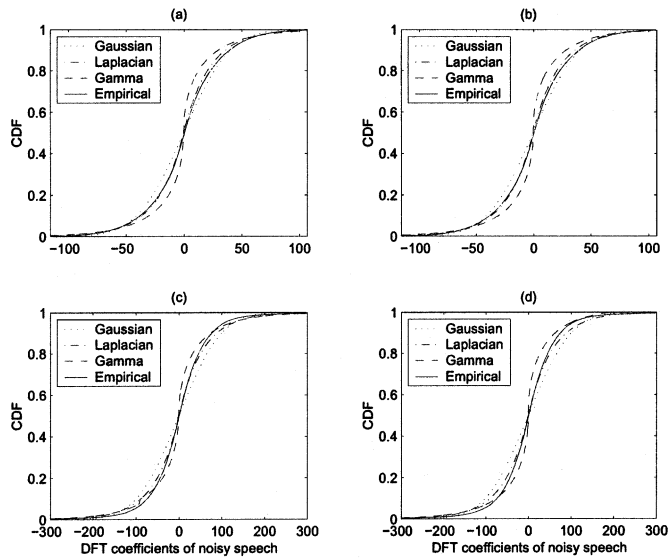


Fig. 2. Approximating the empirical cdf for the DFT coefficients of noisy speech by parametric models (at SNR = 10 dB, white noise) (a) H_0 , real parts of DFT coefficients; (b) H_0 , imaginary parts of DFT coefficients; (c) H_1 , real parts of DFT coefficients; and (d) H_1 , imaginary parts of DFT coefficients.

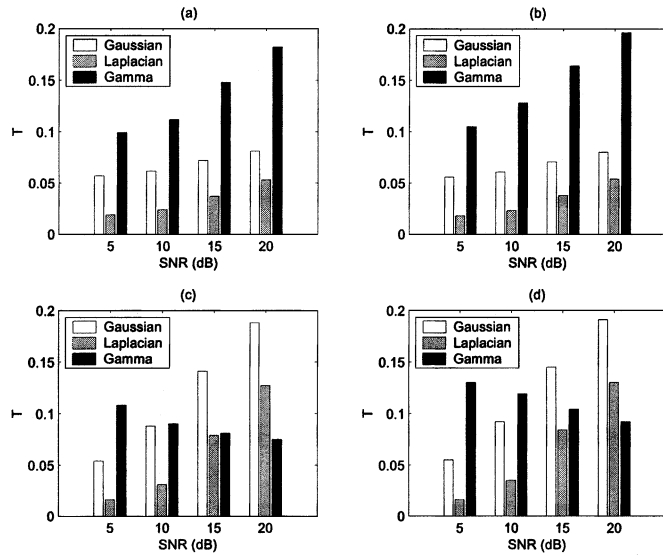


Fig. 3. KS statistics of statistical models for the DFT coefficients of noisy speech under white noise conditions (at SNR = 5, 10, 15, 20 dB) (a) H_0 , real parts of DFT coefficients; (b) H_0 , imaginary parts of DFT coefficients; (c) H_1 , real parts of DFT coefficients; and (d) H_1 , imaginary parts of DFT coefficients.

Fig. 2 shows the empirical cdfs and the assumed parametric distributions under the noise condition at SNR = 10 dB. From Fig. 2, we can see that the Laplacian cdf is closest to the empirical cdf among all the parametric models for both hypotheses H_0 and H_1 . Figs. 3–5 show the results of the KS statistics evaluated under various noise conditions. Summarizing the results, we make the following observations.

- 1) According to Fig. 3, the Laplacian model was best fitted to H_0 at all SNR levels while the next best fit was the Gaussian model. One may ask whether white noise is Gaussian distributed. Again, we see that the DFT coefficients for white noise are better approximated by the Laplacian rather than by others since the frame length

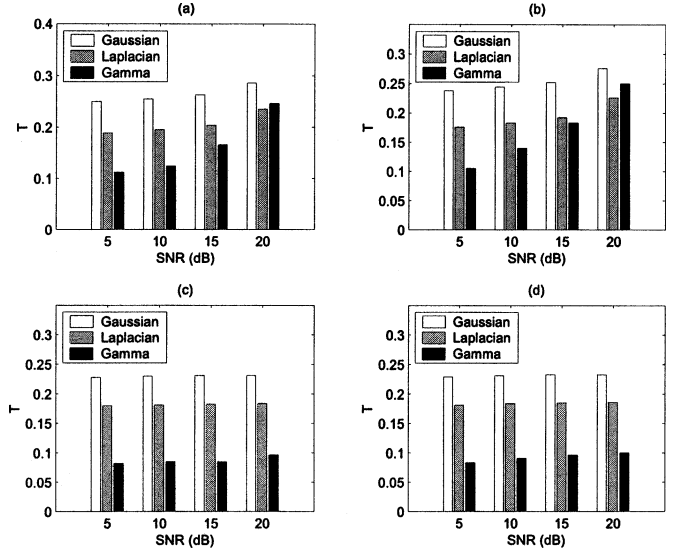


Fig. 4. KS statistics of statistical models for the DFT coefficients of noisy speech under vehicular noise conditions (at SNR = 5, 10, 15, 20 dB) (a) H_0 , real parts of DFT coefficients; (b) H_0 , imaginary parts of DFT coefficients; (c) H_1 , real parts of DFT coefficients; and (d) H_1 , imaginary parts of DFT coefficients.

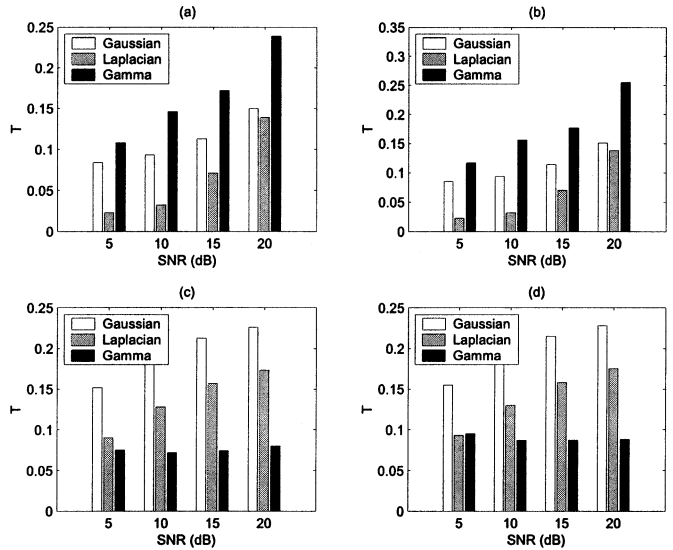


Fig. 5. KS statistics of statistical models for the DFT coefficients of noisy speech under babble noise conditions (at SNR = 5, 10, 15, 20 dB) (a) H_0 , real parts of DFT coefficients; (b) H_0 , imaginary parts of DFT coefficients; (c) H_1 , real parts of DFT coefficients; and (d) H_1 , imaginary parts of DFT coefficients.

is restricted to 10 ms. For large frame lengths (>200 ms), the pdf of DFT coefficients may approach Gaussian, which is not practical for real-time signal processing applications (including VAD) [25].

For the hypothesis H_1 , the Laplacian model was best fitted except for SNR = 20 dB. Specifically, the KS statistic computed from the Gamma pdf became smaller as the SNR increased, while the other models produced larger values. Finally, the KS statistic of the Gamma model was lowest at SNR = 20 dB. For that reason, we may conclude that the Gamma model becomes more suitable as the noisy speech approaches the clean speech spectra under H_1 .

- 2) The KS statistics for the vehicular noise case are shown in Fig. 4. From Fig. 4, it can be seen in every case that the KS statistic of the Gamma model is smaller than those of the Laplacian and Gaussian models with an exception at SNR = 20 dB when H_0 is true. In almost all the cases for the vehicular noise, the Gaussian distribution assumption was found to be a poor choice for characterizing the DFT spectra of noisy speech.
- 3) Fig. 5 shows the KS statistics computed when the babble noise was added to the clean speech waveforms. This time, we note that the results obtained under H_1 show a tendency which is quite different from that produced under H_0 . For H_0 , the KS statistic of the Laplacian model was smallest in almost all the SNR values, with the KS statistic of the Gamma model being always the largest. On the other hand, the Gamma model was found best suitable for the distribution of DFT coefficients of noisy speech when H_1 was true. Also, it should be noted here that the performance of the Gaussian model is worst in almost all conditions. Summarizing the KS test results, we know that it would be beneficial to apply different statistical models for noisy speech DFT spectra depending on the types of noises and SNR levels.

III. LIKELIHOOD RATIO TEST FOR VAD

For VAD based on the assumed statistical models, a decision statistic should be defined for each frequency bin as follows [13]:

$$\Lambda_k \equiv \frac{p(X_k|H_1)}{p(X_k|H_0)} \quad (18)$$

in which Λ_k represents the LR of the k th frequency bin. The decision rule for VAD is established as the geometric mean of the LRs computed for the individual frequency bins such that

$$\log \Lambda = \frac{1}{M} \sum_{k=0}^{M-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (19)$$

with M being the total number of frequency bins and η denoting the threshold of detection. The LR when the Gaussian pdfs are used for speech distribution modeling both for H_0 and H_1 is given by

$$\Lambda_k^G = \frac{p_G(X_k|H_1)}{p_G(X_k|H_0)} \quad (20)$$

$$= \frac{1}{1 + \xi_k} \exp \left\{ \frac{\gamma_k \xi_k}{1 + \xi_k} \right\} \quad (21)$$

where $\xi_k = \lambda_{s,k}/\lambda_{n,k}$ and $\gamma_k = |X_k|^2/\lambda_{n,k}$, and they are called the *a priori* and *a posteriori* SNRs, respectively [8]. On the other hand, in the case of the Laplacian pdfs given by (9) and (10), LR is derived as

$$\Lambda_k^L = \frac{p_L(X_k|H_1)}{p_L(X_k|H_0)} \quad (22)$$

$$= \frac{1}{1 + \xi_k} \exp \left\{ 2 \left(|X_{k(R)}| + |X_{k(I)}| \right) \left(\frac{|\hat{X}_k| - \sqrt{\lambda_{n,k}}}{|\hat{X}_k| \sqrt{\lambda_{n,k}}} \right) \right\} \quad (23)$$

Practically, it is known that instead of applying (23) directly, it is more beneficial to substitute $\sqrt{\lambda_{n,k} + \lambda_{s,k}}$ with the instantaneous magnitude spectrum $|X_k|$ [19]. This approach is referred to as the *instantaneous form*.

Here, we propose an alternative method called the *smoothed form* in which the LR is modified such that

$$\hat{\Lambda}_k^L = \frac{1}{1 + \xi_k} \exp \left\{ 2 \left(|X_{k(R)}| + |X_{k(I)}| \right) \left(\frac{|\hat{X}_k| - \sqrt{\lambda_{n,k}}}{|\hat{X}_k| \sqrt{\lambda_{n,k}}} \right) \right\} \quad (24)$$

where $|\hat{X}_k| = \kappa_L |X_k| + (1 - \kappa_L) \sqrt{\lambda_{n,k} + \lambda_{s,k}}$ and κ_L is a smoothing factor. It is noted that in the smoothed form, a more smooth estimate for the magnitude spectra is applied instead of its instantaneous value. Based on a number of experiments, we have observed that the smoothed form is superior to the instantaneous form. It is worth mentioning that the performance of the LRT largely depends on the selection of the smoothing factor κ_L , and one efficient method to determine κ_L is described in Section V.

For the Gamma pdf, the formulation of the LR is given by

$$\Lambda_k^M = \frac{p_M(X_k|H_1)}{p_M(X_k|H_0)} \quad (25)$$

$$= \frac{1}{\sqrt{1 + \xi_k}} \exp \left\{ \frac{\sqrt{2}}{\sqrt{3}} \left(|X_{k(R)}| + |X_{k(I)}| \right) \cdot \left(\frac{\sqrt{\lambda_{n,k} + \lambda_{s,k}} - \sqrt{\lambda_{n,k}}}{\sqrt{\lambda_{n,k} + \lambda_{s,k}} \sqrt{\lambda_{n,k}}} \right) \right\} \quad (26)$$

Similar to the Laplacian case, we can transform (26) into the corresponding smoothed form as follows:

$$\hat{\Lambda}_k^M = \frac{1}{\sqrt{1 + \xi_k}} \exp \left\{ \frac{\sqrt{2}}{\sqrt{3}} \left(|X_{k(R)}| + |X_{k(I)}| \right) \left(\frac{|\hat{X}_k| - \sqrt{\lambda_{n,k}}}{|\hat{X}_k| \sqrt{\lambda_{n,k}}} \right) \right\} \quad (27)$$

where $|\hat{X}_k| = \kappa_M |X_k| + (1 - \kappa_M) \sqrt{\lambda_{n,k} + \lambda_{s,k}}$ and κ_M is a smoothing parameter for the Gamma model.

IV. SNR ESTIMATION

A robust method to track the noise and speech power spectra and equivalently the SNR is required to complete the VAD algorithm. In this section, we compare several approaches to SNR estimation.

A. Maximum Likelihood (ML) and Decision-Directed (DD) Estimation

In [8], the DD approach is proposed for the estimation of the *a priori* SNR based on a heuristic knowledge. In fact, the DD approach is widely applied to speech enhancement due to its superiority in reducing musical noises. In [14], the DD estimator is applied to VAD to reduce the bias of the ML estimator $\hat{\xi}_k^{\text{ML}} = \gamma_k - 1$ as follows:

$$\hat{\xi}_k^{\text{DD}}(t) = \alpha \frac{|\hat{X}_k(t-1)|^2}{\hat{\lambda}_{n,k}(t-1)} + (1 - \alpha) u[\hat{\gamma}_k(t) - 1] \quad (28)$$

where $u[x] = x$ if $x \geq 0$ and $u[x] = 0$ otherwise; and $\hat{X}_k(t-1)$ is the speech spectral amplitude estimate obtained in the previous frame, for which we use the MMSE estimator. Also, α is a weight which is usually determined in the range (0.95, 0.99) [8]. The DD method provides a smoother estimate of the *a priori* SNR than the ML method, and consequently reduces the undesired fluctuation of the evaluated likelihood ratio during the nonspeech periods. However, one of the drawbacks of the DD method is that the *a priori* SNR appears simply as a delayed version of the *a posteriori* SNR during the active speech periods. Due to such delay, the detection performance of the VAD is significantly degraded at speech onset and offset regions.

B. Predicted (PD) Estimation

To overcome the drawback of the DD method in speech activity periods, we apply the PD approach based on a soft-decision scheme to estimate the *a priori* SNR [17]. Here, we briefly review the PD approach for the estimation of the *a priori* SNR. Let $\{\lambda_{n,k}(t)\}$ and $\{\lambda_{s,k}(t)\}$ denote the power spectra at time t of the background noise and clean speech, respectively. Then, in the PD estimation technique

$$\begin{aligned}\hat{\lambda}_{n,k}(t+1) &= \zeta_n \hat{\lambda}_{n,k}(t) + (1 - \zeta_n) E[|N_k(t)|^2 | X_k(t)] \\ \hat{\lambda}_{s,k}(t+1) &= \zeta_s \hat{\lambda}_{s,k}(t) + (1 - \zeta_s) E[|S_k(t)|^2 | X_k(t)]\end{aligned}\quad (29)$$

where $\hat{\lambda}_{n,k}(t)$, $\hat{\lambda}_{s,k}(t)$ are the estimates for $\lambda_{n,k}(t)$, $\lambda_{s,k}(t)$; and $\zeta_n (= 0.99)$ and $\zeta_s (= 0.98)$ are the parameters for smoothing under a general stationarity assumption on $\mathbf{N}(t)$ and $\mathbf{S}(t)$.⁴

Taking into account the uncertainty for speech absence or presence is possible by incorporating the speech absence probability, which, according to Bayes' rule, is given by [17]

$$p(H_0 | X_k) = \frac{1}{1 + \frac{P(H_1)}{P(H_0)} \Lambda(X_k(t))}. \quad (30)$$

With (30), the expectations in (29) are described as

$$\begin{aligned}E[|N_k(t)|^2 | X_k(t)] &= E[|N_k(t)|^2 | X_k(t), H_0] P(H_0 | X_k(t)) \\ &\quad + E[|N_k(t)|^2 | X_k(t), H_1] P(H_1 | X_k(t))\end{aligned}\quad (31)$$

$$\begin{aligned}E[|S_k(t)|^2 | X_k(t)] &= E[|S_k(t)|^2 | X_k(t), H_0] P(H_0 | X_k(t)) \\ &\quad + E[|S_k(t)|^2 | X_k(t), H_1] P(H_1 | X_k(t))\end{aligned}\quad (32)$$

where

$$\begin{aligned}E[|N_k(t)|^2 | X_k(t), H_0] &= |X_k(t)|^2 \\ E[|S_k(t)|^2 | X_k(t), H_0] &= 0\end{aligned}\quad (33)$$

⁴It seems natural that initial values of $\lambda_{n,k}(t = 0)$ and $\lambda_{s,k}(t = 0)$ are $|X_k(t = 0)|^2$ and zero because we assume that only noise exists on initial several frames, which is similar to the noise suppression rule of enhanced variable rate codec [25]. On the other hand, we employ $\lambda_{s,\min} (= 0.085)$ and $\lambda_{n,\min} (= 0.085)$ instead of zero in terms of $\lambda_{s,k}$ and $\lambda_{n,k}$ for the robust speech quality.

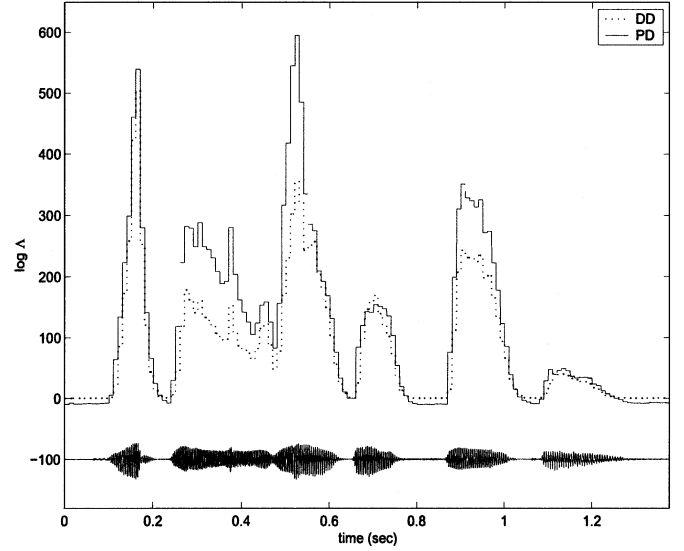


Fig. 6. Comparison of $\log \Lambda$ for the DD and PD estimation with white noise (SNR = 5 dB). Corresponding clean speech waveform is plotted for ease of comparison.

$$\begin{aligned}E[|N_k(t)|^2 | X_k(t), H_1] &= \left(\frac{\hat{\xi}_k^{\text{PD}}(t)}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right) \hat{\lambda}_{n,k}(t) \\ &\quad + \left(\frac{1}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right) |\hat{N}_k(t)|^2\end{aligned}\quad (34)$$

$$\begin{aligned}E[|S_k(t)|^2 | X_k(t), H_1] &= \left(\frac{\hat{\xi}_k^{\text{PD}}(t)}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right) \hat{\lambda}_{s,k}(t) \\ &\quad + \left(\frac{1}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right)^2 |X_k(t)|^2\end{aligned}\quad (35)$$

$$\begin{aligned}E[|N_k(t)|^2 | X_k(t), H_1] &= \left(\frac{1}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right) \hat{\lambda}_{s,k}(t) \\ &\quad + \left(\frac{\hat{\xi}_k^{\text{PD}}(t)}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right) |\hat{S}_k(t)|^2\end{aligned}\quad (36)$$

$$\begin{aligned}E[|S_k(t)|^2 | X_k(t), H_1] &= \left(\frac{1}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right) \hat{\lambda}_{s,k}(t) \\ &\quad + \left(\frac{\hat{\xi}_k^{\text{PD}}(t)}{1 + \hat{\xi}_k^{\text{PD}}(t)} \right)^2 |X_k(t)|^2\end{aligned}\quad (37)$$

with

$$\hat{\xi}_k^{\text{PD}}(t) \equiv \frac{\hat{\lambda}_{s,k}(t)}{\hat{\lambda}_{n,k}(t)} \quad (38)$$

that is the estimate for the *a priori* SNR at time t computed based on the PD approach given by (29). An example of the LR curves obtained by the DD and PD approaches to estimate the *a priori* SNR is displayed in Fig. 6. Due to the delayed characteristics of the DD estimation, the PD estimation approach seems to be able to respond to the SNR variation more rapidly compared

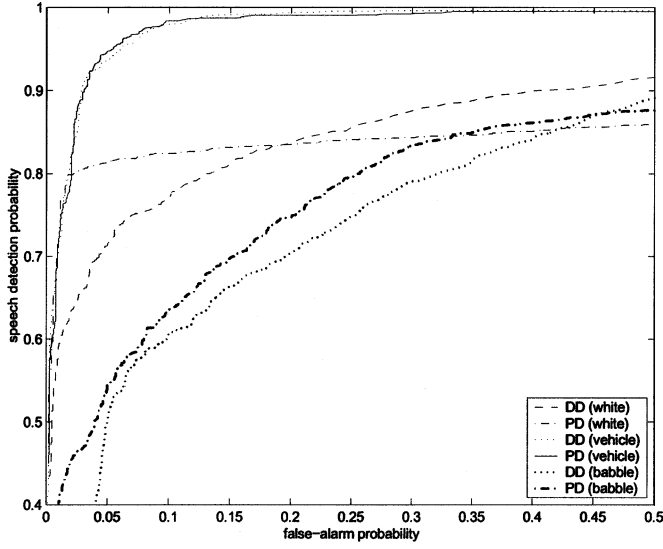


Fig. 7. ROC curves for the DD and PD approaches with the white, vehicular, and babble noise (SNR = 5 dB) for the Gaussian model.

with the DD estimation scheme particularly at the speech onset and offset regions although the PD estimation uses information of the previous frame.

V. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed algorithms, we investigated the speech detection and false-alarm probabilities (P_d and P_f) for each statistical model. We define P_d as the ratio of correct speech decisions to the hand-marked speech frames and P_f as that of false speech decisions to the hand-marked nonspeech frames. To evaluate P_d and P_f , we made reference decisions on clean speech material 64 s long⁵ by labeling manually at every 10 ms frame. The percentage of the hand-marked active speech frames was 54.9%, which consisted of 46.5% voiced sounds and 8.4% unvoiced sounds. For the purpose of making the input data deviate from the pure speech, we added the white, vehicular, and babble noises from the NOISEX-92 database [34] to the clean speech waveforms with varying SNR. The VAD test was carried out for each 10 ms frame.

A. DD and PD Estimation

Both the PD and DD estimation techniques were compared to estimate the *a priori* SNR for each frequency bin. The performance of each approach was measured in terms of the receiver operating characteristic (ROC) curve which shows the tradeoff between P_d and P_f . Fig. 7 shows the ROC curves of the DD and PD estimation approaches for the Gaussian model when the SNR was 5 dB. From Fig. 7, it can be clearly seen that the applied PD estimation technique has performance better than or at least comparable to that of the conventional DD estimation method in most of the noisy conditions except $P_f > 20\%$ ⁶ for the white noise. The difference between the DD and PD is most noticeable under the babble noise. This is mainly due to the fact

⁵See [14], where 46-s-long speech material is used. We used the speech segment that was used in Section II since it seems a desirable way to analyze the performance of the VAD based on the provided statistical analysis.

⁶These regions are not considered in the actual VAD since the false-alarm rate is too high.

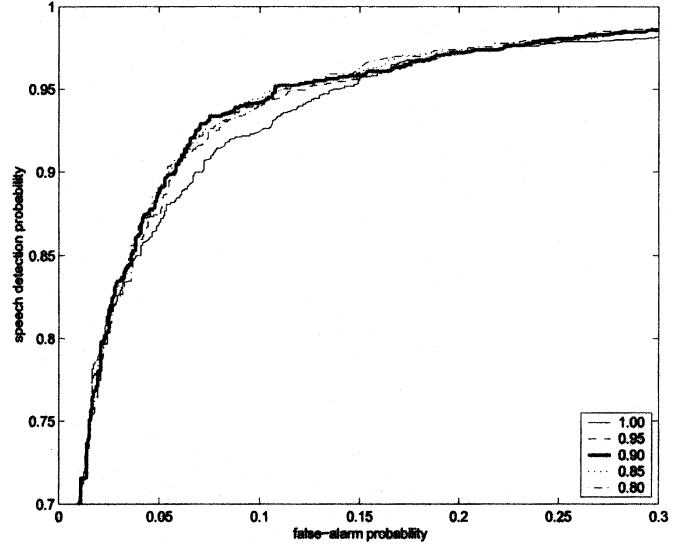


Fig. 8. ROC curves for various κ_L with white noise at 5 dB SNR.

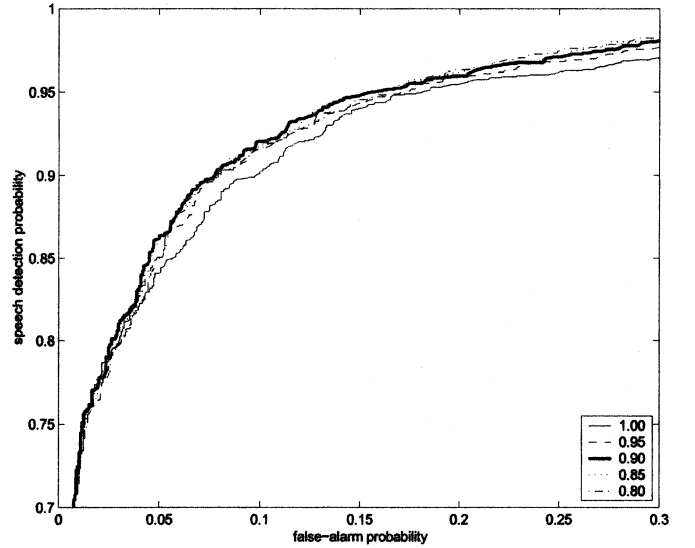


Fig. 9. ROC curves for various κ_M with white noise at 5 dB SNR.

that performance of the PD is more effective in low detection accuracy. From the simulation for the *a priori* SNR estimation, it was found that the performance improvement was mostly due to the rapid SNR tracking capability of the PD technique especially at the speech onset and offset regions.

B. Determination of κ

The choice of the value of parameters κ_L and κ_M is guided by a tradeoff between the instantaneous input signal power and the estimated value. According to several simulation results, we know that it is better to choose the value of κ_L (κ_M) as close to one as possible to increase the detection probability. On the other hand, to decrease the false-alarm probability, we should keep the value of κ_L (κ_M) as close to zero as possible. To find out the optimal κ_L and κ_M , we examined the ROC curves by varying the values of the κ_L (κ_M) based on the aforementioned speech material contaminated by the white noise environment at SNR = 5 dB. Figs. 8 and 9 show the ROC curves generated with different values of the smoothing factor in the Laplacian

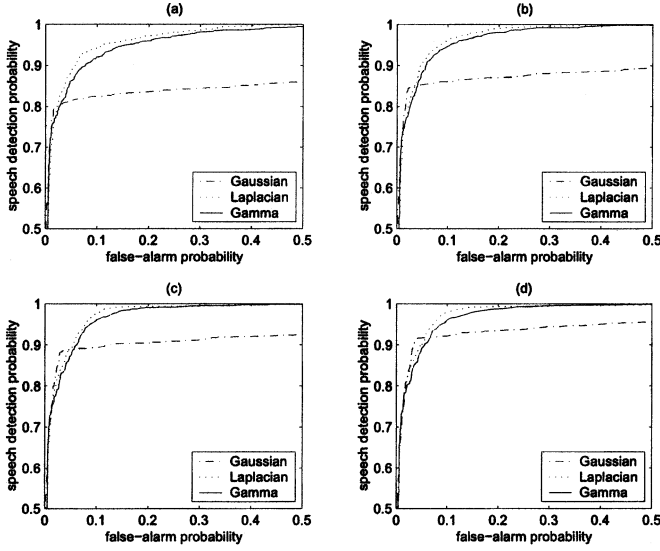


Fig. 10. ROC curves for white noise: (a) 5 dB, (b) 10 dB, (c) 15 dB, and (d) 20 dB.

and Gamma models, respectively. From the results, we can see that $\kappa_L = 0.9$ in the Laplacian model and $\kappa_M = 0.9$ in the Gamma model produced the best detection performance. These two values are confirmed based on the given 64 speech material with aforementioned various noise types and SNRs. In practice, κ_L and κ_M were fixed to 0.9 evaluate the performance of VAD under other noise conditions without an additional tuning in this paper.

The smoothed form offers the choice to increase the performance of the Laplacian and Gamma-based LR, because the smoothed form provides us a smoother estimate for the magnitude spectra compared to the instantaneous value. It seems fair because the Gaussian-based LR also employs the long-term smooth type of $\lambda_{s,k}$ and $\lambda_{n,k}$ in (29), which are adopted in ξ_k seen in (38).

C. Comparison of Overall Performance Based on Statistical Models

In this section, we present the overall performance of the proposed statistical model-based VAD approaches. The ROC curves of the three statistical models are shown in Figs. 10–12 and where the VAD algorithms were applied to the speech data corrupted by the aforementioned noises at a variety of SNRs (5, 10, 15, and 20 dB). VAD algorithms were implemented by combining the PD estimation technique with the smoothed form LRT approach while keeping $\kappa_L = 0.9$ and $\kappa_M = 0.9$. From the results, we could obtain the following observations.

- 1) In the case of the white noise, which is shown in Fig. 10, the Laplacian model-based VAD algorithm outperformed the other approaches. Also, the Gamma model-based one resulted in a better performance than that of the Gaussian model in most of the tested conditions.⁷ Compared with the KS result, a significant difference in ROC curves between proposed two models and the conventional Gaussian model illustrates that not only

⁷See [14], where the all ROC curves of the Gaussian have abrupt turn. We have a similar abrupt pattern for the Gaussian in our work.

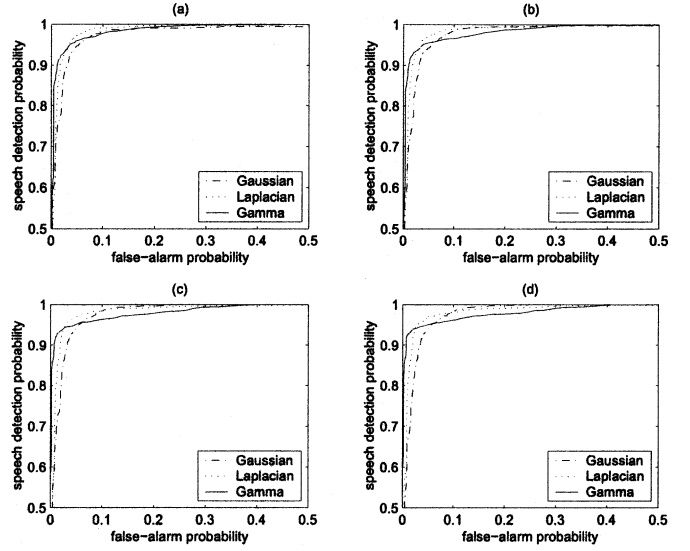


Fig. 11. ROC curves for vehicular noise: (a) 5 dB, (b) 10 dB, (c) 15 dB, and (d) 20 dB.

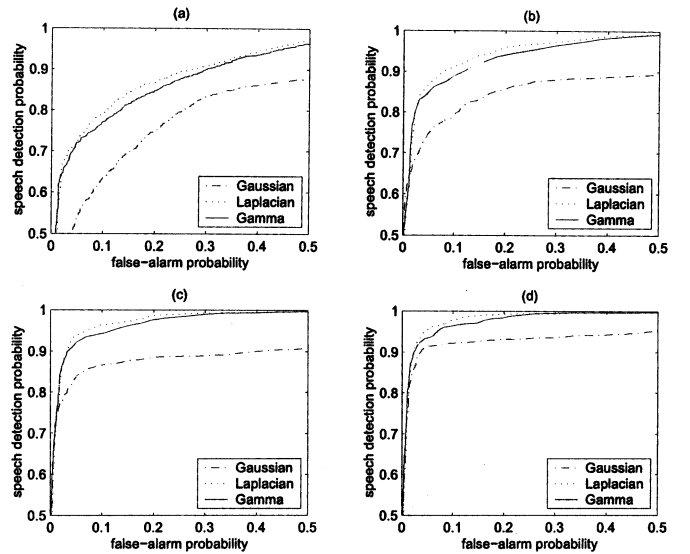


Fig. 12. ROC curves for babble noise: (a) 5 dB, (b) 10 dB, (c) 15 dB, and (d) 20 dB.

the statistically accurate model but also the proposed PD estimation and smoothed form give clearly a better contrast to evaluate the speech activities. On the other hand, as the SNR increased to 20 dB, the Gaussian model yielded a little better performance than the other parametric distributions when $P_f < 0.05$.

- 2) In contrast, the ROC curves look relatively close to each other in the case of the vehicular noise, as can be seen in Fig. 11. However, it is observed that the Gamma model-based VAD algorithm demonstrates a slightly better performance than the other models only when $P_f < 0.02$. It is also noted that the difference in detection accuracy becomes larger as the SNR is increased. Except for the low false-alarm operating condition where $P_f < 0.02$, the Laplacian model-based VAD approach has been found to be the best among the assumed parametric models.

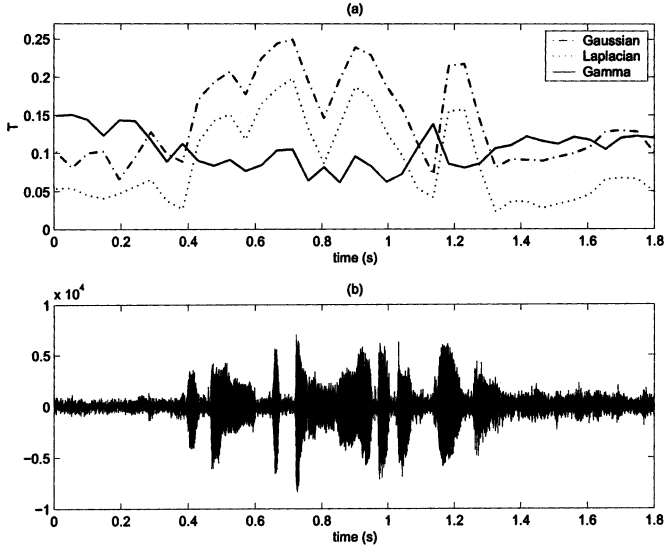


Fig. 13. KS statistics of the three statistical models evaluated for the DFT coefficients (real part) of noisy speech.

On the other hand, the Gaussian-based VAD algorithm outperformed the Laplacian-based method only when $P_f > 0.1$ at SNR = 20 dB.

- 3) From Fig. 12, which shows the results for the babble noise, the VAD algorithm incorporating the Laplacian model yielded a performance superior to both the Gamma and Gaussian pdfs. Moreover, the performance difference became larger as the SNR decreased.

D. Multiple Statistical Models-Based VAD Employing Online KS Test

Until now, the statistical model was selected prior to the VAD process in the presented techniques. From the result of the proposed model-based VAD, we would need a different statistical model under the different hypotheses and types of noise statistics for the performance improvement. Consequently, a robust way is necessary to track the suitable model while keeping the additional computational load below a tolerable level. Specifically, we present the multiple statistical model (MSM)-based VAD algorithm based on online KS test which is effective in dealing with input signal whose statistical properties of the added noise or the SNR change as time flows in a nonstationary manner. We estimate the online KS statistics T using the recently incoming data $\mathbf{X}_m = \{\mathbf{X}(t), \mathbf{X}(t-1), \dots, \mathbf{X}(t-m)\}$ every p frames, where m is the past m previous frames as follows⁸:

$$\hat{T}_t = \lambda_T \hat{T}_{t-p} + (1 - \lambda_T) T_t(\mathbf{X}_m) \quad (39)$$

where λ_T is a smoothing parameter.

Fig. 13 plots the online KS statistics $\hat{T}_t$ for the given statistical models in conjunction with the noisy speech samples. From Fig. 13, we can see that the proposed online KS test can track effectively the actual statistical properties mentioned in Section II. In the case of the babble noise shown in Fig. 13,

⁸ m and p were set to 20 and 10 in the experiment for the good convergence property without a large amount of computational load.

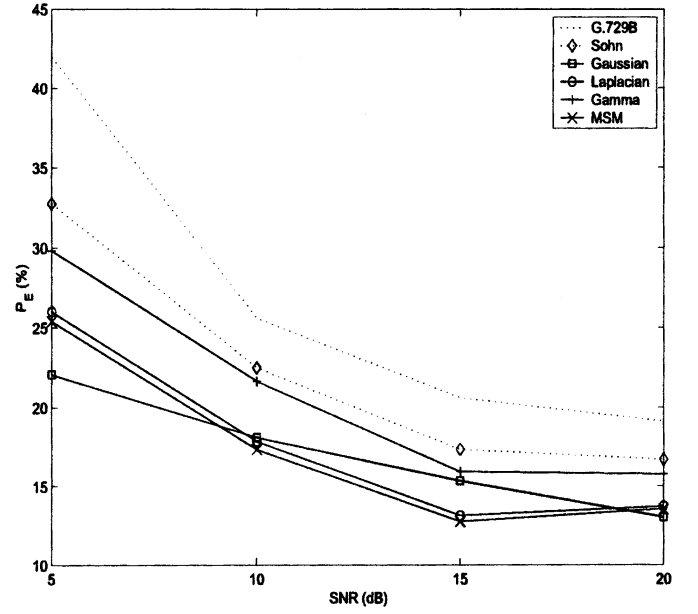


Fig. 14. P_E of the proposed Gaussian, Laplacian, Gamma, MSM, conventional G.729 Annex B, and Sohn's VAD for white noise at various SNR conditions.

the Laplacian performed better than both the Gamma and the Gaussian in terms of the KS statistics during the speech segments. Based on the motivation, the statistical model with the lowest KS statistics prior to VAD is selected every p frames. This method is more realistic, as the statistical model varies according to the spectral distribution of the input signal as time progresses.

E. Speech Detection Error Probability Tests

By examining the detection performance under a variety of noisy conditions, we could select a set of thresholds for each parametric model to apply the LRT. So, we evaluated the performance of the proposed VAD algorithms by fixing the thresholds of the LRT to the selected values and compared the results with those of the ITU-T G.729B [9] and Sohn's algorithm [14]. Specifically, we evaluated the speech detection error probability (P_E), which is the sum of the false alarm and missing probabilities [35]. Figs. 14–16 show the detection results under the white, vehicular, and babble noise, respectively. From Fig. 14, we make the following observations.

- 1) In the case of proposed VAD which uses only Gaussian, Laplacian, and Gamma for the statistical model, we have better performance than the G.729B under almost all evaluated conditions. Also, the Sohn's VAD [14] performed better than G.729B in almost all the testing conditions.
- 2) The results in many conditions demonstrate the superiority of the Laplacian-based VAD compared to the other model-based one and Sohn's method.
- 3) Proposed MSM-based VAD is superior to or at least comparable to the Laplacian-based VAD at all testing conditions. It is evident from the results that the MSM is a more desirable way to characterize the statistical distribution for the DFT coefficients of noisy speech for VAD.

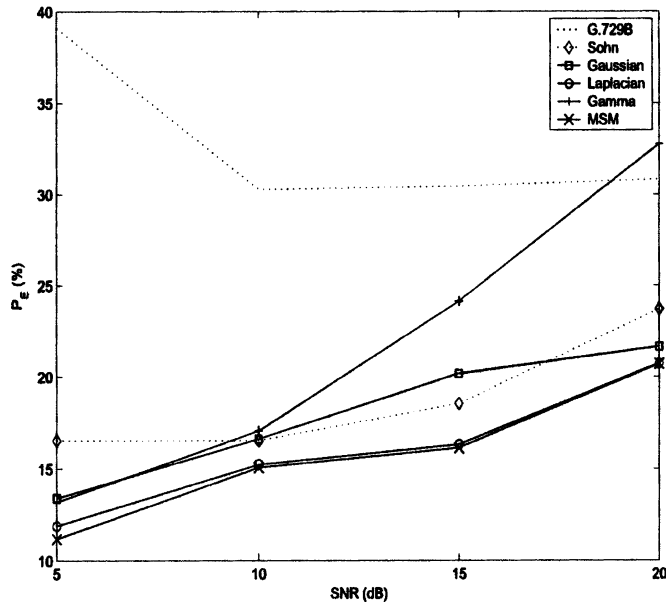


Fig. 15. P_E of the proposed Gaussian, Laplacian, Gamma, MSM, conventional G.729 Annex B, and Sohn's VAD for vehicular noise at various SNR conditions.

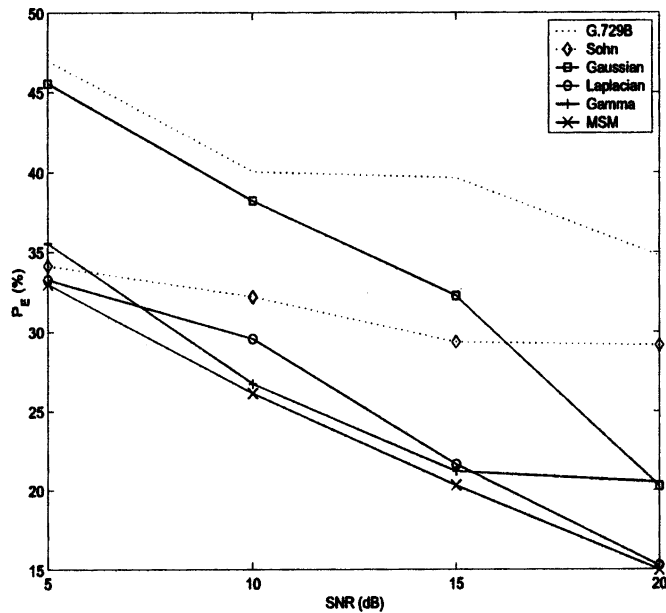


Fig. 16. P_E of the proposed Gaussian, Laplacian, Gamma, MSM, conventional G.729 Annex B, and Sohn's VAD for babble noise at various SNR conditions.

F. Comparison of Computational Load

Even though the superiority of the proposed approaches in terms of objective tests is clear, their computational complexity compared to that of conventional methods should be assessed for a fair comparison. In the case of G.729B, a set of difference parameters is extracted and used for the VAD decision. The parameters are the full band energy, the low band energy, the ZCR, and a spectral measure based on the line spectral frequencies (LSFs). The major differences between the Sohn's algorithm and our approach lie in the fact

that the former provides the different schemes such as the DD estimation and the Hamming window while the latter gives the PD estimation and the trapezoidal window. Also, the noise spectrum adaptation of the Sohn's VAD is quite different from our algorithm.

Although the presented VAD is more intensive with regard to the computational cost when compared to the G.729B VAD since it employs the FFT, they are considered inadequate for comparing the computational load directly since the used feature is combined with the other routines. For example, the LSF is an essential part in the codebook quantization in G.729B, while the FFT is primarily employed in the noise suppression of the many commercial codecs [10], [25].

Table I summarizes the computational load results of the aforementioned six different VAD algorithms. For the input speech data which is 456 s long, the overall computation times (ms) have been calculated and are given.⁹ In the case of external feature, the LSF is used in the G.729B VAD while the FFT is employed in the statistical model-based one. We divide the computation stages into the main and external feature which will give us somewhat clearer comparison. As can be seen, the computation time required for the proposed algorithm turned out to be more than that of the G.729B VAD. The proposed MSM-based VAD is heaviest among the algorithms due to the online KS test while it has a better performance. On the other hand, the computational load of the FFT in our proposed scheme is less than that of the LSF in the G.729B. Totally, the proposed VAD algorithm is about two times as much as the G.729B. But the proposed VAD has an advantage, as it requires only a few parameters to optimize when compared with the G.729B VAD [14], [15].

Even though the computational burden is considerably reduced when we decrease the FFT size, it is not desirable because the detection performance degrades according to the simulation results. Also, it may be interesting to consider the other features such as the mel frequency cepstral coefficients. We expect that the incorporation of a statistical modeling using the features will improve the performance of the VAD. However, in this paper, we have restricted ourselves to the FFT, and the adoption of the other features is a subject of future study.

VI. CONCLUSION

The objective of this paper is to discover the properties of the statistical models for the representation of the noisy speech spectra and to apply the statistical models to VAD. An extensive study and experiments on the statistical models under various noise conditions have made it possible to understand that the complex Laplacian and Gamma pdfs could be strong candidates for a parametric representation of the noisy speech spectra distribution. Based on the proposed statistical models, we have designed the VAD algorithm by means of the LRT with a more robust parameter estimation. Moreover, to improve the detection performance at the speech transition

⁹We used the profiler in the Visual C++ of the Microsoft using Intel Pentium III 933 MHz. Using this, we could check the separate and overall computation time for each function of the reference C code (floating point).

TABLE I
COMPUTATION TIME (ms) USING THE PROFILER IN THE VISUAL C++ OF MICROSOFT. FOR THE EXTERNAL FEATURE,
THE LSF AND THE FFT ARE ADOPTED FOR THE G.729B VAD AND OTHER METHODS, RESPECTIVELY

method	conventional algorithms		proposed algorithms			
	G.729 Annex B	Sohn [14]	Gaussian	Laplacian	Gamma	MSM
VAD	169	3605	3630	3636	3643	4138
External Feature	1974	1110	1002	1002	1002	1002
Total	2143	4715	4632	4638	4645	5140

regions, we have proposed the PD estimation scheme which enables to resolve the drawback of the conventional DD estimation approach. From a number of experiments, it is shown that the performance of the VAD has been significantly improved when we take advantage of the Laplacian model for noisy speech, especially in white and babble noise conditions, while the performance for Gaussian case is worse than the performance of the other cases in most noise conditions. With the incorporation of more accurate statistical model employing the online KS test, an improvement of the detection accuracy was achieved with a little burden of computation load.

ACKNOWLEDGMENT

The authors would like to thank J. Sohn for making his VAD code available and helpful comments and the anonymous reviewers for invaluable suggestions.

REFERENCES

- [1] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using Itakura LPC distance measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1977, pp. 323–326.
- [2] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1994, pp. 237–240.
- [3] J. C. Junqua, B. Reaves, and B. Mark, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Proc. Eurospeech*, 1991, pp. 1371–1374.
- [4] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral feature," in *Proc. IEEE TELCON*, China, 1993, pp. 321–324.
- [5] N. B. Yoma, F. McIness, and M. Jack, "Robust speech pulse-detection using adaptive noise modeling," *Electron. Lett.*, vol. 32, pp. 1350–1352, Jul. 1996.
- [6] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Electr. Eng.*, vol. 139, pp. 377–380, Aug. 1992.
- [7] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 1818–1829, Dec. 1998.
- [8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [9] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70," ITU-T Rec. G. 729, Annex B, 1996.
- [10] 3GPP2, "Selectable mode vocoder service option for wideband spread spectrum communication systems," 3GPP2 C.S0030-0 ver. 1.0, 2001.
- [11] ITU, "Speech coders: Silence compression scheme," ITU-T Rec. G.723.1, Annex A, 1996.
- [12] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels," ETSI EN 301 708 v7.1.1, Dec. 1999.
- [13] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 365–368, 1998.
- [14] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
- [15] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, pp. 276–278, Oct. 2001.
- [16] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, pp. 108–110, May 2000.
- [17] J.-H. Chang and N. S. Kim, "Speech enhancement: New approaches to soft decision," *IEICE Trans. Inf. Syst.*, vol. 27, no. E84-D, pp. 1231–1240, Sep. 2001.
- [18] —, "Speech enhancement using warped discrete cosine transform," in *Proc. IEEE Speech Coding Workshop*, Tsukuba, Japan, Oct. 2002, pp. 175–177.
- [19] —, "Voice activity detection based on complex Laplacian model," *Electron. Lett.*, vol. 39, no. 7, pp. 632–634, Apr. 2003.
- [20] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, pp. 2403–2418, Nov. 2001.
- [21] —, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, pp. 12–15, Jan. 2002.
- [22] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, pp. 113–116, Apr. 2002.
- [23] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 478–482, Jul. 2000.
- [24] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC Residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 217–231, Mar. 2001.
- [25] "Enhanced variable rate codec, speech service option 3 for wideband spectrum digital systems," TIA/EIA/IS-127, 1996.
- [26] D. R. Brillinger, *Time Series: Data Analysis and Theory*. New York: Holden-Day, 1981.
- [27] A. G. Glen, L. M. Leemis, and D. R. Barr, "Order statistics in goodness-of-fit testing," *IEEE Trans. Reliab.*, vol. 50, pp. 209–213, Jun. 2001.
- [28] R. C. Reininger and J. D. Gibson, "Distributions of the two dimensional DCT coefficients for images," *IEEE Trans. Commun.*, vol. COM-31, no. 6, pp. 835–839, Jun. 1983.
- [29] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137–145, Apr. 1980.
- [30] O. Cappé, "Elimination of musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 345–349, Apr. 1994.
- [31] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Orlando, FL, May 2002, pp. 1253–1256.
- [32] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, pp. 204–207, Jul. 2003.
- [33] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.
- [34] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition, II—NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.
- [35] S. M. Kay, *Fundamental of Statistical Signal Processing, Volume 2: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.



Joon-Hyuk Chang (M'03) was born in Daegu, Korea, in 1973. He received B.S. degree in electronics engineering from Kyungpook National University, Daegu, in 1998 and the M.S. and Ph.D. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 2000 and 2004, respectively.

From March 2000 to April 2005, he was with Netdus Corp., Seoul, as a Chief Engineer. From May 2004 to April 2005, he was with the University of California, Santa Barbara, in a postdoctoral position to work on adaptive signal processing and audio coding. In May 2005, he joined Korea Institute of Science and Technology, Seoul, as a Research Scientist to work on speech recognition. Currently, he is full time Lecturer with the Faculty of Electronic Engineering, Inha University, Incheon, Korea. His research interests are in speech coding, speech enhancement, speech recognition, audio coding, and adaptive signal processing.



Nam Soo Kim (M'98) received the B.S. degree in electronics engineering from Seoul National University (SNU), Seoul, Korea, in 1988 and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology in 1990 and 1994, respectively.

From 1994 to 1998, he was with Samsung Advanced Institute of Technology as a Senior Member of Technical Staff. Since 1998, he has been with the School of Electrical Engineering, SNU, where he is currently an Associate Professor. His research area

includes speech signal processing, speech recognition, speech/audio coding, speech synthesis, adaptive signal processing, machine learning, and mobile communication.



Sanjit K. Mitra (S'59–M'63–SM'69–F'74–LF'01) received the B.Sc. degree (with honors) in physics and the M.Sc.(Tech.) degree in radio physics and electronics in India in 1953 and 1956, respectively. He received the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1960 and 1962, respectively.

Since 1977, he has been a Professor of electrical and computer engineering at the University of California, Santa Barbara, where he was Chair of the department from 1979 to 1982. He has published

more than 600 papers in the areas of analog and digital signal processing and image processing. He has authored or coauthored 12 books and has received five patents. He has held visiting appointments in Australia, Austria, Brazil, Finland, India, Japan, Singapore, and the United Kingdom.

Dr. Mitra is a Fellow of AAAS and SPIE. He is a member of the European Association for Signal Processing (EURASIP), the U.S. National Academy of Engineering, and the Norwegian Academy of Technological Sciences and an Academician of the Academy of Finland. He is a Corresponding Member of the Croatian Academy of Sciences and Arts and the Academy of Engineering of Mexico. He received the 1973 F. E. Terman Award and the 1985 AT&T Foundation Award from the American Society of Engineering Education, the 1989 Education Award and the 2000 Mac Van Valkenburg Society Award from the IEEE Circuits and Systems Society, the Distinguished Senior U.S. Scientist Award from the Alexander von Humboldt Foundation of Germany in 1989, the 1996 Technical Achievement Award and the 2001 Society Award from the IEEE Signal Processing Society, the IEEE Millennium Medal in 2000, the McGraw-Hill/Jacob Millman Award from the IEEE Education Society in 2001, and the 2002 Technical Achievement Award from EURASIP. He is the corecipient of the 2000 Blumlein-Browne-Willans Premium of the Institution of Electrical Engineers (London) and the 2001 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award. He has received Honorary Doctorate degrees from the Tampere University of Technology, Tampere, Finland, and the Technical University of Bucharest, Bucharest, Romania. He has served IEEE in various capacities, including as President of the IEEE Circuits and Systems Society in 1986.