Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 16921, 10 pages
doi:10.1155/2007/16921

Research Article

Robust In-Car Speech Recognition Based on Nonlinear Multiple Regressions

Weifeng Li,[1] Kazuya Takeda,[1] and Fumitada Itakura[2]

[1] *Graduate School of Information Science, Nagoya University, Nagoya 464-8603, Japan*
[2] *Department of Information Engineering, Faculty of Science and Technology, Meijo University, Nagoya 468-8502, Japan*

Received 31 January 2006; Revised 10 August 2006; Accepted 29 October 2006

Recommended by S. Parthasarathy

We address issues for improving handsfree speech recognition performance in different car environments using a single distant microphone. In this paper, we propose a nonlinear multiple-regression-based enhancement method for in-car speech recognition. In order to develop a data-driven in-car recognition system, we develop an effective algorithm for adapting the regression parameters to different driving conditions. We also devise the model compensation scheme by synthesizing the training data using the optimal regression parameters and by selecting the optimal HMM for the test speech. Based on isolated word recognition experiments conducted in 15 real car environments, the proposed adaptive regression approach shows an advantage in average relative word error rate (WER) reductions of 52.5% and 14.8%, compared to original noisy speech and ETSI advanced front end, respectively.

1. INTRODUCTION

The mismatch between training and testing conditions is one of the most challenging and important problems in automatic speech recognition (ASR). This mismatch may be caused by a number of factors, such as background noise, speaker variation, a change in speaking styles, channel effects, and so on. State-of-the-art ASR techniques for removing the mismatch usually fall into the following three categories [1]: robust features, speech enhancement, and model compensation. The first approach seeks parameterizations that are fundamentally immune to noise. The most widely used speech recognition features are the Mel-frequency cepstral coefficients (MFCCs) [2]. MFCC's lack of robustness in noisy or mismatched conditions has led many researchers to investigate robust variants or novel feature extraction algorithm. Some of these researches could be perceptually based on, for example, the PLP [3] and RASTA [4], while other approaches are related to the auditory processing, for example, gammatone filter [5] and EIH model [6].

Speech enhancement approach aims to perform noise reduction by transforming noisy speech (or feature) into an estimate that more closely resembles clean speech (or feature). Examples falling in this approach include spectral subtraction [7], Wiener filter, cepstral mean normalization (CMN) [8], codeword-dependent cesptral normalization (CDCN) [9], and so on. Spectral subtraction was originally proposed in the context of the enhancement of speech quality, but it can be used as a preprocessing step for recognition. However, its performance suffers from the annoying "musical tone" artifacts. CMN performs the simple linear transformation and aims to remove the cepstral bias. Although effective for the convolutional distortions, this technique is not successful for the additive noise. CDCN may be somewhat intensive to compute since it depends on the online estimation of the channel and additive noise through an iterative EM approach. Model compensation approach aims to adapt or transform acoustic models to match the noisy speech feature in a new testing environment. The representative methods include multistyle training [8], maximum-likelihood linear regression (MLLR) [10], and Jacobian adaptation [11, 12]. Their main disadvantage is that they require the retraining of a recognizer or adaptation data, which leads to much higher complexity than speech enhancement approach. Most speech enhancement and model compensation methods are accomplished by linear functions such as simple bias removal, affine transformation, linear regression, and so on. However, it is well known that distortion caused even by additive noise only is highly nonlinear in the log-spectral or cepstral domain. Therefore, a nonlinear transformation or compensation is more appropriate.

The use of a neural network allows us to automatically learn the nonlinear mapping functions between the reference and testing environments. Such a network can handle additive noise, reverberation, channel mismatches, and combinations of these. Neural-network-based feature enhancement has been used in conjunction with a speech recognizer. For example, Sorensen used a multilayer network for noise reduction in the isolated word recognition under F-16 jet noise [13]. Yuk and Flanagan employed neural networks to perform telephone speech recognition [14]. However, the feature enhancement they implemented was performed in the ceptral domain and the clean features were estimated using the noisy features only.

In previous work, we proposed a new and effective multimicrophone speech enhancement approach based on multiple regressions of log spectra [15] that used multiple spatially distributed microphones. Their idea is to approximate the log spectra of a close-talking microphone by effectively the combining of the log spectra of distant microphones. In this paper, we extend the idea to single-microphone case and propose that the log spectra of clean speech are approximated through the nonlinear regressions of the log spectra of the observed noisy speech and the estimated noise using a multilayer perceptron (MLP) neural network. Our neural-network-based feature enhancement method incorporates the noise information and can be viewed as a generalized log spectral subtraction.

In order to develop a data-driven in-car recognition system, we develop an effective algorithm for adapting the regression parameters to different driving conditions. In order to further reduce the mismatch between training and testing conditions, we synthesize the training data using the optimal regression parameters, and train multiple hidden Markov models (HMMs) over the synthesized data. We also develop several HMM selection strategies. The devised system results in a universal in-car speech recognition framework including both the speech enhancement and the model compensation.

The organization of this paper is as follows: in Section 2, we describe the in-car speech corpus used in this paper. In Section 3, we present the regression-based feature enhancement algorithm, and the experimental evaluations are outlined in Section 4. In Section 5, we present the environmental adaptation and model compensation algorithms. Then the performance evaluation on the adaptive regression-based speech recognition framework is reported in Section 6. Finally Section 7 concludes this paper.

2. IN-CAR SPEECH DATA AND SPEECH ANALYSIS

A data collection vehicle (DCV) has been specially designed for developing the in-car speech corpus at the Center for Integrated Acoustic Information Research (CIAIR), Nagoya University, Nagoya, Japan [16]. The driver wears a headset with a close-talking microphone (#1 in Figure 1) placed in it. Five spatially distributed microphones (#3 to #7) are placed around the driver. Among them, microphone #6, located at the visor location to the speaker (driver), is the closest to the speaker. The speech recorded at this microphone (also named "visor mic.") is used for speech recognition in this paper. A four-element linear microphone array (#9 to #12) with an interelement spacing of 5 cm is located at the visor position.

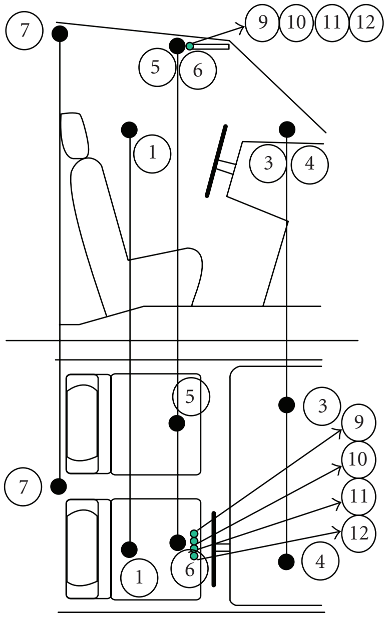

Figure 1: Side view (top) and top view (bottom) of the arrangement of multiple spatially distributed microphones and the linear array in the data collection vehicle.

The test data includes Japanese 50 word sets under 15 driving conditions (3 driving environments ×5 in-car states = 15 driving conditions as listed in Table 1). Table 2 shows the average signal-to-noise ratio (SNR) for each driving condition. For each driving condition, 50 words are uttered by each of 18 speakers. A total of 7000 phonetically balanced sentences (uttered by 202 male speakers and 91 female speakers) were recorded for acoustical modeling. (3600 of them were collected in the idling-normal condition and 3400 of them were collected while driving the DCV on the streets near Nagoya University (city-normal condition).)

Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz. For spectral analysis, a 24-channel MFB analysis is performed on 25-millisecond-long windowed speech, with a frame shift of 10 milliseconds. Spectral components lower than 250 Hz are filtered out to compensate for the spectrum of the engine noise, which is concentrated in the lower-frequency region. Log MFB parameters are then estimated. The estimated log MFB vectors are transformed into 12 mean normalized Mel-frequency cepstral coefficients (CMN-MFCC) using discrete cosine transformation (DCT) and mean normalization, after which the time derivatives (Δ CMN-MFCC) are calculated.

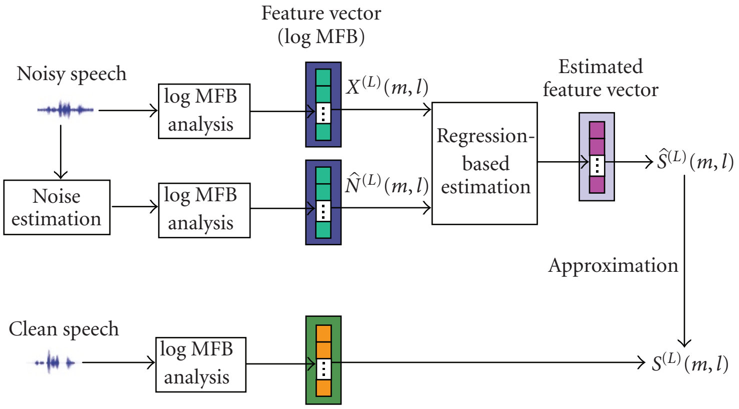

FIGURE 2: Concept of regression-based feature enhancement.

TABLE 1: Fifteen driving conditions (3 driving environments ×5 in-car states).

Driving environments	In-car states
Idling "i" City driving "c" Expressway driving "e"	Normal "n" CD player on "s" Air conditioner (AC) on at low level "l" Air conditioner (AC) on at high level "h" Window (near the driver) open "w"

TABLE 2: The average SNR values (dB) for 15 driving conditions ("i-n" indicates the idling-normal condition, and so on).

Cond.	SNR	Cond.	SNR	Cond.	SNR
i-n	13.41	c-n	9.58	e-n	7.24
i-s	8.82	c-s	8.13	e-s	7.16
i-l	9.56	c-l	8.92	e-l	7.30
i-h	6.84	c-h	6.49	e-h	5.92
i-w	8.87	c-w	6.55	e-w	4.29

3. ALGORITHMS

3.1. Regression-based feature enhancement

Let $s(i)$, $n(i)$, and $x(i)$, respectively, denote the reference clean speech (referred to the speech at the close-talking microphone in this paper), noise, and observed noisy signals. By applying a window function and analysis using short-time discrete Fourier transform (DFT), in the time-frequency domain we have $S(k,l)$, $\hat{N}(k,l)$, and $X(k,l)$, where k and l denote frequency bin and frame indexes, respectively. The hat above N denotes the estimated version. After the Mel-filter-bank (MFB) analysis and the log operation, we obtain $S^{(L)}(m,l)$, $X^{(L)}(m,l)$, and $\hat{N}^{(L)}(m,l)$, that is,

$$\begin{aligned} S^{(L)}(m,l) &= \log \sum_k r_{m,k}\,|S(k,l)|, \\ X^{(L)}(m,l) &= \log \sum_k r_{m,k}\,|X(k,l)|, \\ \hat{N}^{(L)}(m,l) &= \log \sum_k r_{m,k}\,|\hat{N}(k,l)|, \end{aligned} \tag{1}$$

where $r_{m,k}$ denotes the weights of the mth filter bank. The idea of the regression-based enhancement is to approximate $S^{(L)}(m,l)$ with the combination of $X^{(L)}(m,l)$ and $\hat{N}^{(L)}(m,l)$, as shown in Figure 2. Let $\hat{S}^{(L)}(m,l)$ denote the estimated log MFB ouput of the mth filter bank at frame l, and it can be obtained from the inputs of $X^{(L)}(m,l)$ and $\hat{N}^{(L)}(m,l)$. In particular, $\hat{S}^{(L)}(m,l)$ can be obtained using the linear regression, that is,

$$\hat{S}^{(L)}(m,l) = b_m + w_m^{(x)} X^{(L)}(m,l) + w_m^{(n)} \hat{N}^{(L)}(m,l), \tag{2}$$

where the parameters $\Theta = \{b_m, w_m^{(x)}, w_m^{(n)}\}$ are obtained by minimizing the mean-squared error:

$$\mathcal{E}(m) = \sum_{l=1}^{L} \left[S^{(L)}(m,l) - \hat{S}^{(L)}(m,l)\right]^2, \tag{3}$$

over the training examples. Here, L denotes the number of training examples (frames).

On the other hand, $\hat{S}^{(L)}(m,l)$ can be obtained by applying multilayer perceptron (MLP) regression method, where a network with one hidden layer composed of 8 neurons is used,[1] that is,

$$\begin{aligned} &\hat{S}^{(L)}(m,l) \\ &\quad = f\left(X^{(L)}, \hat{N}^{(L)}\right) \\ &\quad = b_m + \sum_{p=1}^{8} \left(w_{m,p} \tanh\left(b_{m,p} + w_{m,p}^{(x)} X^{(L)} + w_{m,p}^{(n)} \hat{N}^{(L)}\right)\right), \end{aligned} \tag{4}$$

[1] The network was determined experimentally.

where the filter bank index m and the index frame l are dropped for compactness. $\tanh(\cdot)$ is the tangent hyperbolic activation function. The parameters $\Theta=\{b_m, w_{m,p}, w_{m,p}^{(x)}, w_{m,p}^{(n)}, b_{m,p}\}$ are found by minimizing (3) through the back-propagation algorithm [17].

The proposed approach is cast into single-channel methodology because once the optimal regression parameters are obtained by regression learning, they can be utilized in the test phase, where the speech of the close-talking microphone is no longer required. Multiple regressions mean that regression is performed for each Mel-filter bank. The use of minimum mean-squared error (MMSE) in the log spectral domain is motivated by the fact that log spectral measure is more related to the subjective quality of speech [18] and that some better results have been reported with log distortion measures [19].[2]

Although neural networks have been employed for feature enhancement (e.g., [13, 14]) in cepstral domain, the input used for the estimation of the clean feature in their algorithms is the noisy feature only. The proposed method incorporates the noise information through the noise estimation, and can be viewed as a generalized log spectral subtraction. In this paper, $|\hat{N}(k,l)|$ is estimated using the two-stage noise spectra estimator proposed in [20]. Based on our previous studies, the incorporation of the noise information contributed a significant performance gain of about 3% absolute improvement in recognition accuracies, compared to that using the noisy feature only.

3.2. Comparison with the spectral subtraction

The *spectral subtraction* (SS) [7] is a simple but effective technique for cleaning the speech from the additive noise. It was originally developed for the speech quality enhancement. However, they may also serve as a preprocessing step for the speech recognition. Let the corrupted speech signal $x(i)$ be represented as

$$x(i) = s(i) + n(i), \tag{5}$$

where $s(i)$ is the clean speech signal and $n(i)$ is the noise signal. By applying a window function and the analysis using short-time discrete Fourier transform (DFT), we have

$$X(k,l) = S(k,l) + N(k,l), \tag{6}$$

where k and l denote frequency bin and frame indexes, respectively. For compactness, we will drop both k and l. Assuming that the clean speech s and the noise n are statistically independent, the power spectrum of clean speech $|S|^2$ can be estimated as

$$|\hat{S}|^2 = |X|^2 - |\hat{N}|^2, \tag{7}$$

where $|\hat{N}|^2$ is the estimated noise power spectrum. To reduce the annoying "musical tone" artifacts, SS can be modified as [21]

$$|\hat{S}|^2 = \begin{cases} |X|^2 - \alpha|\hat{N}|^2 & \text{if } |X|^2 > \beta|\hat{N}|^2, \\ \beta|\hat{N}|^2 & \text{otherwise,} \end{cases} \tag{8}$$

by introducing the subtraction factor α and the spectral flooring parameter β. SS can be also implemented in the amplitude domain and the subband domain [22].

Although the proposed regression-based method and SS are implemented in the different domains, both of them estimate the features of the clean speech using those of noisy speech and estimated noise. In (8), the SS method results in a simple subtraction of the weighted noise power spectra from the noisy speech power spectra. In most literatures, the parameters α and β are usually determined experimentally. Compared with SS, the regression-based method employs more general nonlinear models, and can benefit from the regression parameters, which are statistically optimized. Moreover, the proposed method makes no assumption about the independence of speech and noise, and can deal with more complicated distortions rather than the additive noise only.

3.3. Comparison with the log-spectra amplitude (LSA) estimator

The *log-spectra amplitude* (LSA) estimator [23], proposed by Ephraim and Malah, also employs minimum mean-squared errors (MMSEs) cost function in the log domain. However, this approach explicitly assumes a Gaussian distribution for the clean speech and the additive noise spectra. Under this assumption, by using the MMSE estimation on log-spectral amplitude, we can obtain the estimated amplitude of clean speech as

$$|\hat{S}| = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt\right)\cdot|X|, \tag{9}$$

where the *a priori* and *a posteriori* SNRs are defined by $\xi = E\{|S|^2\}/E\{|\hat{N}|^2\}$ and $\gamma = E\{|X|^2\}/E\{|\hat{N}|^2\}$, respectively, where $E\{\cdot\}$ denotes the expectation operator. ν is defined by

$$\nu = \frac{\xi}{1+\xi}\gamma. \tag{10}$$

To reduce he "musical tone" artifacts, the dominant parameter, the *a priori* SNR ξ, is calculated using the smoothing technique, that is, the "decision-directed" method [24].

Compared to SS method, the LSA estimator results in a nonlinear model and is well known for its reduction of the "musical tone" artifacts [25]. However, the LSA estimator is based on the additive noise model and Gaussian distributions of speech and noise spectra, which is not true for realistic data [26]. In the LSA estimator, the dominant parameter ξ is simply estimated by the smoothing over the neighbor frames, and the smoothing parameter is usually determined experimentally. On the contrary, the proposed

[2] In [19], Porter and Boll found that for speech recognition, minimizing the mean-squared errors in the log |DFT| is superior to using all other DFT functions and to spectral magnitude subtraction.

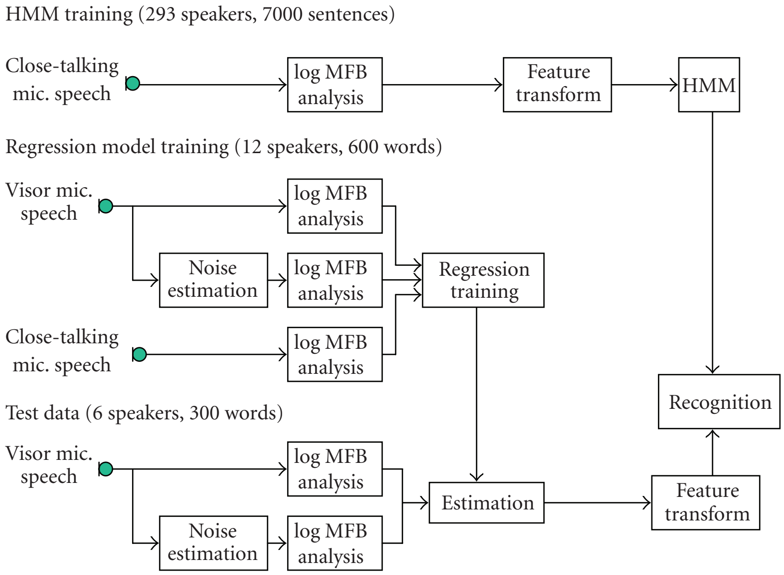

Figure 3: Diagram of regression-based speech recognition for a particular driving condition.

method makes no assumptions regarding the additive noise model, nor about the Gaussian distributions of speech and noise spectra. All the regression parameters in the proposed regression method are obtained through the statistical optimization.

4. REGRESSION-BASED SPEECH RECOGNITION EXPERIMENTS

4.1. Experimental setup

We performed isolated word recognition experiments on the 50 word sets under 15 driving conditions as listed in Table 1. In this section, we assume that the driving conditions are known as a priori, and the regression parameters are trained for each condition. For each driving condition, the data uttered by 12 speakers (6 males and 6 females) is used for learning the regression models, and the remaining words uttered by 6 speakers (3 males and 3 females) are used for recognition. A diagram of the in-car regression-based speech recognition for a particular driving condition is given in Figure 3. The structure of the hidden Markov models (HMMs) used in this paper is fixed, that is,

(1) three-state triphones based on 43 phonemes that share 1000 states;
(2) each state has 32-component mixture Gaussian distributions;
(3) the feature vector is a 25-dimensional vector (12CMN-MFCC+12Δ CMN-MFCC + Δ log energy).[3]

For comparison, we performed the following experiments:

original: recognition of the original noisy speech (#6 in Figure 1) speech using the corresponding HMM;

SS: recognition of the speech enhanced using the spectral subtraction (SS) method with (8);

LSA: recognition of the speech enhanced using the log-spectra amplitude (LSA) estimator;

linear regression: recognition of the speech enhanced using the linear regression with (2);

nonlinear regression: recognition of the speech enhanced using the nonlinear regression with (4).

Note that the acoustic models, used for the "SS," "LSA," and the regression method, are trained over the speech at the close-talking microphone (#1 in Figure 1).

4.2. Speech recognition results

The recognition performance averaged over the 15 driving conditions is given in Figure 4. From this figure, it is found that all enhancement methods are effective and outperform the original noisy speech. The linear regression method obtains a higher recognition accuracy than the spectral subtraction method. We contribute it to the statistical optimization of the regression parameters in the linear regression method. The LSA estimator outperforms the linear regression method for its highly nonlinear estimation. The best recognition performance is achieved by the nonlinear regression method for its more flexible model and statistical optimization of the regression parameters. The superiority of the nonlinear regression method is also confirmed by the subjective and objective evaluation experiments on the quality of the enhanced

[3] The regression is also performed on the log energy parameter. The estimated log MFB and the log energy outputs are first converted into CMN-MFCC vectors using DCT and mean normalization. Then the derivatives are calculated.

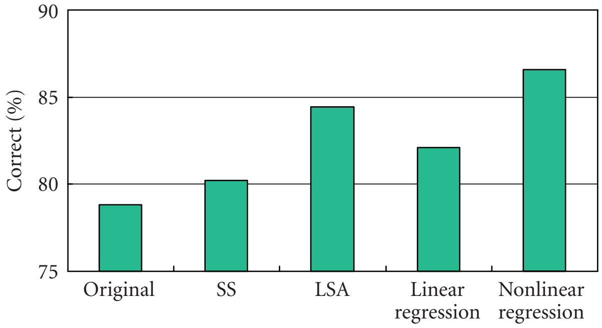

Figure 4: Recognition performance of different speech enhancement methods (averaged over 15 driving conditions).

speech [27].[4] Therefore, the nonlinear regression method is used in the following experiments.

5. ENVIRONMENTAL ADAPTATION AND MODEL COMPENSATION

5.1. Adaptive enhancement of an input speech signal

In the regression-based recognition systems described above, each driving condition was assumed to be known as a prior information and the regression parameters were trained within each driving condition. To develop a data-driven in-car recognition system, regression weights should be adapted automatically to different driving conditions. In this section, we discriminate in-car environments by using the information of the nonspeech signals. In our experiments, Mel-frequency cepstral coefficients (MFCCs) are selected for the environmental discrimination because of their good discriminating ability, even in audio classification (e.g., [28, 29]). The MFCC features are extracted frame by frame from nonspeech signals (preceding the utterance by 200 milliseconds, i.e., 20 frames), their means in one noisy signal are computed, and they are then concatenated into a feature vector:

$$\mathbf{R} = [\bar{c}_1, \ldots, \bar{c}_{12}, \bar{e}], \tag{11}$$

where c_i and e denote ith-order MFCC and log energy, respectively. The upper bar denotes the mean values of the features. Since the variances among the elements in $\mathbf{R}$ are different, each element is normalized so that their mean and variance are 0 and 1, respectively. The prototypes of the noise clusters are obtained by applying the *K-means-clustering* algorithm [30] to the feature vectors extracted from the training set of the nonspeech signals.

The basic procedure of the proposed method is as follows. (1) Cluster the noise signals (i.e., short-time nonspeech segments preceding the utterances) into several groups. (2) For each noise group, train optimal regression weights using the speech segments. (3) For unknown input speech, find a corresponding noise group using the nonspeech segments and perform the estimation with the optimal weights of the selected noise group, that is, the log MFB outputs of clean speech can be estimated by

$$\widehat{\mathbf{S}}^{(L)} = f_k(\mathbf{X}^{(L)}, \widehat{\mathbf{N}}^{(L)}), \tag{12}$$

where $\mathbf{X}^{(L)}$ and $\widehat{\mathbf{N}}^{(L)}$ indicate the log MFB vector obtained from noisy speech and estimated noise, respectively. $f_k(\cdot)$ corresponds to the nonlinear mapping function in Section 3.1, where the cluster ID k is specified by minimizing the Euclidian distance between $\mathbf{R}$ and the centroid vectors.

In our experiments, the vectors $\mathbf{R}'s$, exacted from the first 20-frame nonspeech part of the signals by 12 speakers, are used to cluster the noise conditions, and those by another six speakers are used for testing, as shown in Figure 5.

5.2. Regression-based HMM training

In our previous work [27], we generated the enhanced speech signals, by performing the regressions in the log spectral domain (for each frequency bin). Though few "musical tone" artifacts were found in the regression-enhanced signals compared to those obtained using spectral subtraction-based methods, some noise still remained in the regression-enhanced signals. We believe there will exist a mismatch between training and testing conditions, if we use HMM trained over clean data to test the regression-enhanced speech. In order to reduce the mismatch and incorporate the statistical characteristics of the test conditions, we adopt the K sets of optimal weights obtained from each clustered group to synthesize 7000-sentence training data, that is, we simulated $7000 \times K$ sentences based on K clustered noise environments. Then K HMMs are trained over each of the synthesized 7000-sentence training data, as shown in Figure 5.

5.3. HMM selection

For the recognition of an input speech signal x, an HMM is selected from K HMMs based on the following two strategies.

(1) ID-based strategy

This strategy tries to select an HMM trained over the simulated training data, which are close to the test noise environment, that is,

$$\widehat{H}(x) = \sum_{k=1}^{K} \delta(D(x), D(H_k))H_k, \tag{13}$$

where the *Kronecker delta* function $\delta(\cdot,\cdot)$, has value 1 if its two arguments match, and value 0 otherwise [30]. $D(x)$ and $D(H_k)$ denote the cluster ID of an input signal x and of the kth HMM H_k, respectively.

[4] In our previous work [27], we generated the enhanced speech signals by performing the regressions in the log spectral domain (for each frequency bin).

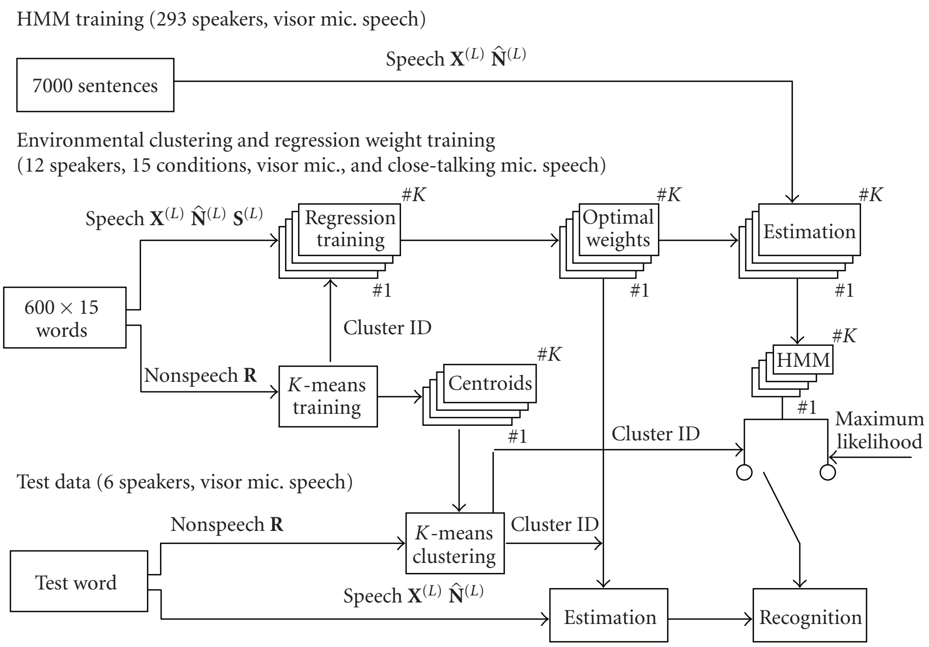
HMM training (293 speakers, visor mic. speech)
7000 sentences
Speech $\mathbf{X}^{(L)}$ $\hat{\mathbf{N}}^{(L)}$
Environmental clustering and regression weight training
(12 speakers, 15 conditions, visor mic., and close-talking mic. speech)
Speech $\mathbf{X}^{(L)}$ $\hat{\mathbf{N}}^{(L)}$ $\mathbf{S}^{(L)}$
Regression training
#K
#1
Optimal weights
#K
#1
Estimation
#K
#1
600 × 15 words
Cluster ID
Nonspeech R
K-means training
Centroids
#K
#1
HMM
#K
#1
Maximum likelihood
Cluster ID
Test data (6 speakers, visor mic. speech)
Nonspeech R
K-means clustering
Cluster ID
Test word
Speech $\mathbf{X}^{(L)}$ $\hat{\mathbf{N}}^{(L)}$
Estimation
Recognition

Figure 5: Diagram of adaptive regression-based speech recognition. $\mathbf{X}^{(L)}$, $\hat{\mathbf{N}}^{(L)}$, and $\mathbf{S}^{(L)}$ denote the log MFB outputs obtained from observed noisy speech, estimated noise, and reference clean speech, respectively. **R** denotes the vector representation of the driving environment using (11).

(2) Maximum-likelihood- (ML-) based strategy

This strategy tries to select the HMM that outputs maximum likelihood (likelihood selection [31]), that is,

$$\hat{H}(x) = \arg\max_{H} \{P(x \mid H_1), \dots, P(x \mid H_K)\}, \qquad (14)$$

where $P(x \mid H_k)$ indicates the log likelihood of an input signal x by using the kth HMM H_k.

5.4. *Analysis of the proposed framework*

There are some common points in the stereo-based piecewise linear compensation for environments (SPLICE) method [32, 33] and our feature enhancement in Section 5.1. Both of them are stereo-based and consist of two steps: finding the optimal "codeword" and performing the codeword-dependent compensation (see (12)). However, the proposed enhancement method does not need any Gaussian assumption required in SPLICE and turns out to be a general nonlinear compensation. Synthesizing the training data using the optimal regression weights obtained in the test environments is similar to training data contaminations [1], but the proposed one incorporates the information of test environments implicitly. Regression-based HMM training and HMM selection can be viewed as a kind of nonlinear model compensation, which can incorporate the information of the testing environments. A combination of feature enhancement and HMM selection results in a universal speech recognition framework where both the noisy features and the acoustic models are compensated.

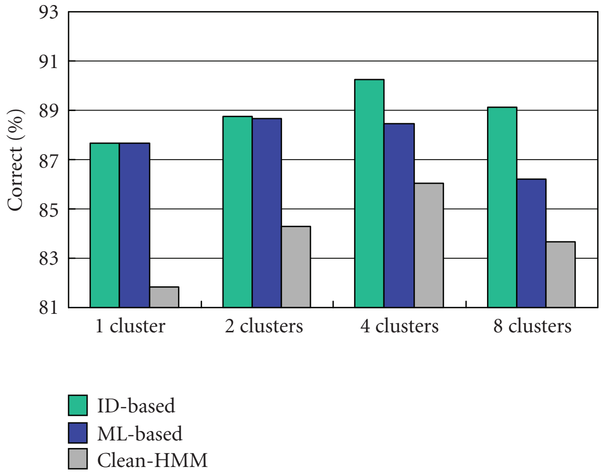

Figure 6: Recognition performance for different clusters using adaptive regression methods (averaged over 15 driving conditions).

6. PERFORMANCE EVALUATION

Figure 6 shows the word recognition accuracies for different numbers of clusters using adaptive regression methods. It is found that the recognition performance is improved significantly by using adaptive regression methods compared to those of "clean-HMM," which is trained over the speech at the close-talking microphone. As the number of clusters increases up to four, the recognition accuracies increase consistently due to there being more noise (environmental) information available. However, too many clusters (e.g., eight or

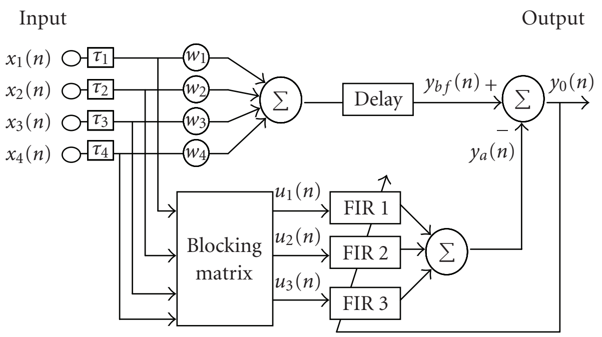

FIGURE 7: Block diagram of generalized sidelobe canceller.

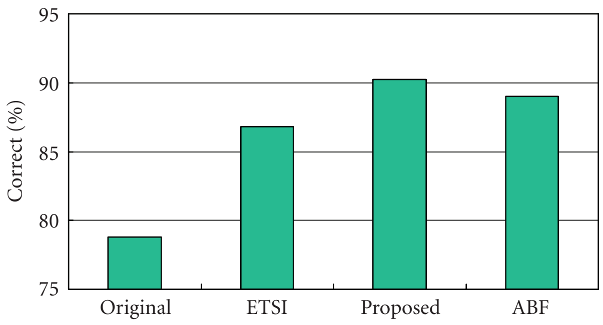

FIGURE 8: Recognition performance of different speech enhancement methods (averaged over 15 driving conditions).

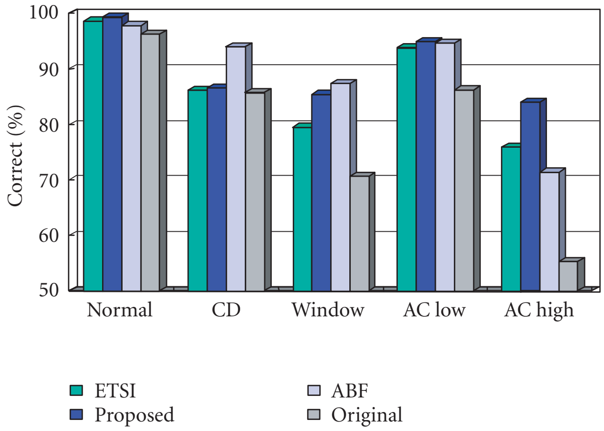

FIGURE 9: Recognition performance for five in-car states by using different methods. Each group represents one in-car state listed in Table 1. Within each group, the bars represent the recognition accuracy by using different methods: ETSI-ETSI advanced front end; proposed—the best performance in Figure 6; ABF-adaptive beamformer; original—recognition of the original noisy speech (no processing).

above) yield a degradation of the recognition performance. Although the two adaptive regression-based recognition systems perform almost identically in the two-cluster case, "ID-based" yields a more stable recognition performance across the numbers of clusters, and the best recognition performance is achieved using "ID-based" and with four clusters.

For comparison, we also performed recognition experiments based on the ETSI advanced front end [34], and an adaptive beamformer (ABF). The acoustic models used for the ETSI advanced front end and the adaptive beamforming were trained over the training data they processed. For the adaptive beamformer, the *generalized sidelobe canceller* (GSC) [35] is applied to our in-car speech recognition. Four linearly spaced microphones (#9 to #12 in Figure 1) with an interelement spacing of 5 cm at the visor position are used. The architecture of the GSC used is shown in Figure 7. In our experiments, τ_i is set equal to zero since the speakers (drivers) sit directly in front of the array line, while w_i is set equal to 1/4. The delay is chosen as half of the adaptive filter order to ensure that the component in the middle of each of the adaptive filters at time n corresponds to $y_{bf}(n)$. The blocking matrix takes the difference between the signals at the adjacent microphones. The three FIR filters are adapted sample by sample using the normalized least-mean square (NLMS) method [36].

Figure 8 shows the recognition performance averaged over the 15 driving conditions. "original" cites from Figure 4 and "proposed" cites the best recognition performance achieved in Figure 6. It is found that all the enhancement methods outperform the original noisy speech. Recalling Figure 4, ETSI advanced front end yields higher recognition accuracy than the LSA estimator. The proposed method significantly outperforms ETSI advanced front end and even performs better than adaptive beamforming, which uses as many as four microphones. Recalling Figure 6, it is found that the regression-based method with even one cluster outperforms ETSI advanced front end. This clearly demonstrates the superiority of the adaptive regression method.

We also investigated the recognition performance averaged over five in-car states as listed in Table 1. The results are shown in Figure 9. It is found that the adaptive regression method outperforms ETSI advanced front end in all the five in-car states, especially when AC is on at high level and when the window near the driver is open. Adaptive beamforming is very effective when the CD player is on and when the window near the driver is open. This suggests that adaptive beamforming with multiple microphones can suppress the noise coming from undesired directions quite well due to its spatial filtering capability. However, in the remaining three in-car states (diffuse noise cases), it does not work as well as the adaptive regression method. Because the proposed method is based on statistical optimization and the present noise estimation cannot track the rapidly changing nonstationary noise, it can be found from this figure that the proposed method works rather well under the stationary noise (e.g., air conditioner on), but has some problems in the nonstationary noise (e.g., CD player on).

7. CONCLUSIONS

In this paper, we have proposed a nonlinear multiple-regression-based feature enhancement method for in-car speech recognition. In the proposed method, the log Mel-filter-bank (MFB) outputs of clean speech are approximated through the nonlinear regressions of those obtained from the noisy speech and the estimated noise. The proposed feature enhancement method incorporates the noise estimation and can be viewed as generalized log-spectral subtraction. Compared with the spectral subtraction and the log-spectral amplitude estimator, the proposed one statistically optimizes the model parameters and can deal with more complicated distortions.

In order to develop a data-driven in-car recognition system, we have developed an effective algorithm for adapting the regression parameters to different driving conditions. We also devised the model compensation scheme by synthesizing the training data using the optimal regression parameters and by selecting the optimal HMM for the test speech. The devised system turns out to be a robust in-car speech recognition framework, in which both feature enhancement and model compensation are performed. The superiority of the proposed system was demonstrated by a significant improvement in recognition performance in the isolated word recognition experiments conducted in 15 real car environments.

In Section 5, a hard decision is made for environmental selection. However, when the system encounters a new noise type, a soft or fuzzy logic decision is desirable, and should be one of future work. The present speech recognition system has not addressed the problem of interference by rapidly changing nonstationary noise. For example, our experiments confirmed that the present recognition system did not work well when CD player was on. In the nonstationary noise cases, the accuracy of noise estimation is very important in successful applications of denoising schemes. Some recursive noise estimation algorithm such as "iterated extended Kalman filter" [37] may be helpful for our speech recognition system.

ACKNOWLEDGMENT

This work is partially supported by a Grant-in-Aid for Scientific Research (A) (15200014).

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[5] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, John Wiley & Sons, New York, NY, USA, 1999.

[6] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, 1994.

[7] 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.

[8] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing—A Guide to Theory, Algorithm, and System Development*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2001.

[9] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pa, USA, 1990.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[11] S. Sagayama, Y. Yamaguchi, and S. Takahashi, "Jacobian adaptation of noisy speech models," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 396–403, Santa Barbara, Calif, USA, December 1997.

[12] R. Sarikaya and J. H. L. Hansen, "Improved Jacobian adaptation for fast acoustic model adaptation in noisy speech recognition," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, pp. 702–705, Beijing, China, October 2000.

[13] H. B. D. Sorensen, "A cepstral noise reduction multi-layer neural network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '91)*, vol. 2, pp. 933–936, Toronto, Ontario, Canada, May 1991.

[14] D. Yuk and J. Flanagan, "Telephone speech recognition using neural networks and hidden Markov models," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99)*, vol. 1, pp. 157–160, Phoenix, Ariz, USA, March 1999.

[15] W. Li, K. Takeda, and F. Itakura, "Adaptive log-spectral regression for in-car speech recognition using multiple distributed microphones," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 340–343, 2005.

[16] N. Kawaguchi, S. Matsubara, H. Iwa, et al., "Construction of speech corpus in moving car environment," in *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP '00)*, pp. 362–365, Beijing, China, October 2000.

[17] S. Haykin, *Neural Networks—A Comprehensive Foundation*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1999.

[18] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[19] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '84)*, vol. 2, pp. 18A.2.1–18A.2.4, San Diego, Calif, USA, 1984.

[20] W. Li, K. Itou, K. Takeda, and F. Itakura, "Two-stage noise spectra estimation and regression based in-car speech recognition using single distant microphone," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. I, pp. 533–536, Philadelphia, Pa, USA, March 2005.

[21] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP '79), vol. 4, pp. 208–211, Washington, DC, USA, April 1979.

[22] J. Chen, K. K. Paliwal, and S. Nakamura, "Sub-band based additive noise removal for robust speech recognition," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 571–574, Aalborg, Denmark, September 2001.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error-log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[25] O. Cappe and J. Laroche, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 84–93, 1995.

[26] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 1, pp. 253–256, Orlando, Fla, USA, May 2002.

[27] W. Li, K. Itou, K. Takeda, and F. Itakura, "Subjective and objective quality assessment of regression-enhanced speech in real car environments," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 2093–2096, Lisbon, Portugal, September 2005.

[28] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99)*, vol. 1, pp. 149–152, Phoenix, Ariz, USA, March 1999.

[29] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 2, pp. 1941–1944, Orlando, Fla, USA, May 2002.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[31] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multimicrophone," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 3, pp. 1747–1750, Istanbul, Turkey, June 2000.

[32] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP '00)*, pp. 806–809, Beijing, China, October 2000.

[33] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 217–220, Aalborg, Denmark, September 2001.

[34] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced frontend feature extraction algorithm; compression algorithm," ETSI ES 202 050 v1.1.1, 2002.

[35] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[36] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2002.

[37] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1995.

Weifeng Li received the B.E. degree in mechanical electronics at Tianjin University, China, in 1997. He received the M.E. and Ph.D. degrees in information electronics at Nagoya University, Japan, in 2003 and 2006. Currently, he is a Research Scientist at the IDIAP Research Institute, Switzerland. His research interests are in the areas of machine learning, speech signal processing, and robust speech recognition. He is a Member of the IEEE.

Kazuya Takeda received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1983, 1985, and 1994, respectively. In 1986, he joined Advanced Telecommunication Research Laboratories (ATR), where he was involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R&D Laboratories and participated in a project for constructing voice-activated telephone extension system. He has joined Graduate School of Nagoya University in 1995. Since 2003, he has been a Professor at Graduate School of Information Science at Nagoya University. He is a Member of the IEICE, IEEE, and the ASJ.

Fumitada Itakura earned undergraduate and graduate degrees at Nagoya University. In 1968, he joined NTT's Electrical Communication Laboratory in Musashino, Tokyo. He completed his Ph.D. in speech processing in 1972. He worked on isolated word recognition at Bell Labs from 1973 to 1975. In 1981, he was appointed as Chief of the Speech and Acoustics Research Section at NTT. In 1984, he took a professorship at Nagoya University. After 20 years, he retired from Nagoya University and joined Meijo University in Nagoya. His major contributions include theoretical advances involving the application of stationary stochastic process, linear prediction, and maximum likelihood classification to speech recognition. He patented the PARCOR vocoder in 1969 the LSP in 1977. His awards include the IEEE ASSP Senior Award, 1975, an award from Japan's Ministry of Science and Technology, 1977, the 1986 Morris N. Liebmann Award (with B. S. Atal), the 1997 IEEE Signal Processing Society Award, and the IEEE third millennium medal. He is a Fellow of the IEEE, a Fellow of the IEICE, and a Member of the ASJ.