A Novel Mask Estimation Method Employing Posterior-Based Representative Mean Estimate for Missing-Feature Speech Recognition

Wooil Kim, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—This paper proposes a novel mask estimation method for missing-feature reconstruction to improve speech recognition performance in various types of background noise conditions. A conventional mask estimation method based on spectral subtraction degrades performance, due to incorrect estimation of the noise signal which fails to accurately represent the variations of background noise during the incoming speech utterance. The proposed mask estimation method utilizes a Posterior-based *Representative* Mean (PRM) estimate for determining the reliability of the input speech spectral components, which is obtained as a weighted sum of the mean parameters of the speech model using the posterior probability. To obtain the noise-corrupted speech model, a model combination method is employed, which was proposed in our previous study for a feature compensation method. Experimental results demonstrate that the proposed mask estimation method provides more separable distributions for the reliable/unreliable component classifier compared to the conventional mask estimation method. The recognition performance is evaluated using the Aurora 2.0 framework over various types of background noise conditions and the CU-Move real-life in-vehicle corpus. The performance evaluation shows that the proposed mask estimation method is considerably more effective at increasing speech recognition performance in various types of background noise conditions, compared to the conventional mask estimation method which is based on spectral subtraction. By employing the proposed PRM-based mask estimation for missing-feature reconstruction, we obtain +23.41% and +9.45% average relative improvements in word error rate for all four types of noise conditions and CU-Move corpus, respectively, compared to conventional mask estimation methods.**

***Index Terms*—Background noise, mask estimation, missing-feature, posterior-based representative mean (PRM) estimate, robust speech recognition.**

I. Introduction

Acoustic environment mismatch between training and operating conditions for actual speech recognition systems severely degrades recognition performance, with background noise as one of the primary corrupting sources. Typical examples can be found in the corpora of NOISEX-92 [3], Speechdat-Car [4], SPINE (SPeech In Noise Environments, [5]), UTDrive [6], CU-Move [7], the National Gallery of Spoken Word (NGSW) [8], Collaborative Digitization Program (CDP) [9], Speech Under Simulated and Actual Stress (SUSAS) including Lombard effect [10], and others, which make speech recognition technology challenging in real-life scenarios. To minimize this mismatch, extensive research has been conducted in recent decades, which includes many types of speech/feature enhancement methods such as spectral subtraction, cepstral mean normalization, and a variety of feature compensation schemes [2], [11]–[18]. Various model adaptation techniques have been successfully employed such as the maximum *a posteriori* (MAP), maximum-likelihood linear regression (MLLR), and parallel model combination (PMC) [19]–[21]. Recently, missing-feature methods have shown promising results [22]–[29].

In this paper, the missing-feature method is considered as a solution to address background noise for speech recognition. This method depends primarily on characteristics of speech that are resistant to noise, rather than on the characteristics of the noise itself, showing its effectiveness at improving speech recognition in adverse environments [22], [23], [25]. The missing-feature method consists of two steps. The first step is estimation of a "mask" which determines which spectral parts of the noisy input speech are unreliable. The second step is to reconstruct the unreliable regions or bypass them for alternative processing.

This paper focuses on the step of mask estimation. One of the most common conventional methods for mask estimation employs the signal-to-noise (SNR) ratio, where the noise signal is estimated from non-speech segments and the clean speech signal is obtained by applying spectral subtraction method [23], [25], [27], [30]. This SNR-based mask estimation method generally depends on the performance of spectral subtraction. Since the noise estimate would not effectively represent the change of background noise during the actual speech utterance, it could provide incorrect estimation of clean speech by spectral subtraction, resulting in performance degradation of mask estimation.

In previous studies, Bayesian classifier based mask estimation methods have been proposed [31]–[33], where several robust speech features were employed, and artificially-generated noise samples were used for training the classifier for the purpose of an environment-independent mask estimation method. However, their performance combined with missing-feature reconstruction method was still outperformed by other conventional preprocessing methods for robust speech recognition. A method to evaluate the spectral reliability using the likelihood

Manuscript received January 23, 2010; revised August 29, 2010; accepted October 23, 2010. Date of publication December 13, 2010; date of current version May 13, 2011. This project was supported in part by the Air Force Research Laboratory (AFRL) through a subcontract to RADC, Inc., under FA8750-09-C-0067 (Approved for public release. Distribution unlimited), and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. A preliminary portion of this study was presented at the IEEE ASRU-2009 workshop [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Gales.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: wikim@utdallas.edu; john.hansen@utdallas.edu).

Digital Object Identifier 10.1109/TASL.2010.2091633

computed from a hidden Markov model (HMM) has also been proposed [34]. A number of studies on mask estimation exploiting spatial information from multiple microphones of incoming speech have also been conducted [35]–[37]; however, they are beyond our focus in this paper where we are interested only in single channel input.

In this paper, a novel mask estimation method for missing-feature reconstruction is proposed to improve speech recognition in background noise conditions. The proposed method utilizes the *representative* mean estimates of clean speech and noise-corrupted speech which are obtained using the posterior probability. A model combination method is employed to generate the noise-corrupted speech model for the proposed mask estimation method. It will be demonstrated that the proposed posterior-based representative mean estimates provide more reliable mask estimation, by decreasing the risk of incorrect estimates of clean speech which is observed in the spectral subtraction based method. The proposed mask estimation method, combined with the missing-feature reconstruction method, will be evaluated on various types of background noise conditions including car, factory, speech babble, and background music, and also the CU-Move in-vehicle data.

This paper is organized as follows. We first review a conventional mask estimation method in Section II. Section III presents details of the proposed mask estimation method, followed by cluster-based missing-feature reconstruction in Section IV Representative experimental procedures and their results are presented with discussion in Section V. Finally, in Section VI we state the main conclusions of our work.

II. Conventional Mask Estimation Method Based on Spectral Subtraction

In this paper, we consider a conventional mask estimation method which employs spectral subtraction to estimate clean speech [23], [30]. In this method, an averaged spectrum of the noise signal $\widetilde{n}^{\{ls\}}(m)$ at the mth frequency band in the log-spectral domain is estimated from silence (i.e., non-speech) segments, which are assumed to exist at the beginning and ending parts of the input speech in this study. Here, $\{ls\}$ indicates the log-spectral domain. These log-spectral coefficients are obtained by taking a logarithm of the Mel-filterbank outputs which are generated during a standard Mel-frequency cepstral coefficients (MFCCs) feature extraction.

In this spectral subtraction based mask estimation method, a subtraction of the estimated speech $\widetilde{x}^{\{ls\}}(t,m)$ obtained by spectral subtraction from the input noise-corrupted speech $y^{\{ls\}}(t,m)$ is compared to a threshold as follows:

$$y^{\{ls\}}(t,m) - \widetilde{x}^{\{ls\}}(t,m) \underset{\text{reliable}}{\overset{\text{unreliable}}{\gtreqless}} \zeta_{\text{SS}} \tag{1}$$

where

$$\widetilde{x}^{\{ls\}}(t,m) = \begin{cases} \log\{\exp(y^{\{ls\}}(t,m)) - \exp(\widetilde{n}^{\{ls\}}(m))\}, \\ \qquad\qquad\qquad \text{if } y^{\{ls\}}(t,m) > \widetilde{n}^{\{ls\}}(m) \\ \beta\, y^{\{ls\}}(t,m), \quad \text{otherwise.} \end{cases} \tag{2}$$

Here, the threshold ζ_{SS} is empirically determined in our experiment and β is a flooring factor.

This conventional mask estimation method mostly relies on the estimated clean speech signal, the correctness of which is dependent on the performance of spectral subtraction as given by (2). Since in general the noise estimate is obtained from silence segments, the estimated noise signal does not represent the temporal variations of the noise signal within the speech utterance, especially for time-varying background noise conditions, resulting in incorrect estimation of clean speech signal by spectral subtraction. Therefore, the mask estimation method based on spectral subtraction would degrade in performance for time-varying background noise conditions[1].

In our initial study [1], we also evaluated another type of mask estimation as a conventional method, which employs signal-to-noise ratio, which also generally depends on noise estimates from silence segments in a manner equivalent to the spectral subtraction-based method; however, the spectral subtraction based method with (1) and (2) showed consistently better performance compared to the SNR-based method. Furthermore, considering that the proposed mask estimation method in this study employs statistical estimates of input noisy speech and clean speech, we believe that the spectral subtraction method is a more "comparable" conventional method which uses input noisy speech and estimated clean speech.

III. Mask Estimation Employing Posterior-Based Representative Mean Estimate

To address the performance degradation of the spectral subtraction-based mask estimation due to incorrect estimation of background noise and clean speech signal, we propose to use estimates of model parameters for the reliability decision, and not directly use estimates of the noise and clean speech. In this paper, we present a new mask estimation method utilizing a *representative* mean estimate for measuring the reliability of spectral components of the input speech, which is determined by posterior probability. Sections IV–VI present the entire procedure of the proposed mask estimation method step by step.

A. Step 1: Speech Model Estimation Employing Model Combination

In our previous study, we proposed the Parallel Combined Gaussian Mixture Model (PCGMM)-based feature compensation method, showing improved speech recognition performance in various types of background noise conditions [2]. In this method, the noise-corrupted speech model (i.e., Gaussian mixture model, GMM) is generated by combining the clean speech GMM and noise model. A series of experiments in this study has confirmed that the noise-corrupted speech model obtained by the PCGMM procedure effectively characterizes the input noise-corrupted speech. From this motivation, we integrate the PCGMM-based model estimation method for obtaining the speech model into our mask estimation method in this study. As presented in Step 2, by employing the PCGMM-based model estimation, an advantage emerges that enables us to calculate the representative mean estimate for the

[1]Here "time-varying" background noise does not only include non-stationary noise (e.g., speech babble, background music) but also slowly time-varying background noise which is widely considered to be stationary noise such as car noise condition.

clean speech by using the same posterior probability of input noise-corrupted speech.

The distribution of the clean speech feature X in the cepstral domain is represented with a GMM consisting of K components as follows:

$$p(X) = \sum_{k=1}^{K} \omega_k \mathcal{N}(X; \boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k}). \tag{3}$$

A noise model is estimated from silence (i.e., non-speech) segments within the input speech as a single Gaussian pdf $(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ in the cepstral domain. The noise-corrupted speech model is obtained through a model combination procedure using the clean speech and noise models, which was employed by the PCGMM-based feature compensation method [2] as

$$(\boldsymbol{\mu}_{Y,k}, \boldsymbol{\Sigma}_{Y,k}) = \mathcal{F}[(\boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k}), (\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)] \tag{4}$$

where $\mathcal{F}[\cdot]$ denotes a function representing the model combination. In this study, we employ "log-normal approximation" method for the model combination, where it is assumed that the addition of two log-normal distributions also results in a log-normal formulation [2], [21].

Before combining the clean speech and noise models, it is required to convert the model parameters from the cepstral domain to the log-spectral domain. The mean and covariance of the cepstral domain are transformed to the log-spectral domain using an inverse discrete cosine transform (DCT) as follows:

$$\begin{aligned} \boldsymbol{\mu}^{\{ls\}} &= \mathbf{C}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\Sigma}^{\{ls\}} &= \mathbf{C}^{-1}\boldsymbol{\Sigma}(\mathbf{C}^{-1})^T. \end{aligned} \tag{5}$$

After both models for clean speech and noise are converted into the log-spectral domain by (5), the model parameters of the noisy speech distribution can be estimated using the model combination procedure which is implied by (4). Finally, the parameters of the noisy speech model must be returned to the cepstral domain via the DCT transform, which is the inverse process of (5). The resulting GMM of the noise-corrupted speech is represented in the cepstral domain as follows:

$$p(Y) = \sum_{k=1}^{K} \omega_k \mathcal{N}(Y; \boldsymbol{\mu}_{Y,k}, \boldsymbol{\Sigma}_{Y,k}) \tag{6}$$

where the same weight ω_k is just used as the clean speech model in (3) as carried over to (6).

B. Step 2: Posterior-Based Representative Mean Estimation

In the proposed mask estimation method, *posterior-based representative mean* (PRM) estimates of noise-corrupted and clean speech at time t are employed for determining the reliability of the spectral component. Here, the PRM estimate of the noise-corrupted speech $\widetilde{\boldsymbol{\mu}}_Y(t)$ in the cepstral domain is defined as a weighted sum of mean parameters of the noise-corrupted speech $\boldsymbol{\mu}_{Y,k}$, using the posterior probability $p(k|Y(t))$ as shown as

$$\widetilde{\boldsymbol{\mu}}_Y(t) = \sum_{k=1}^{K} p(k|Y(t))\boldsymbol{\mu}_{Y,k}. \tag{7}$$

The posterior probability $p(k|Y(t))$ in (7) is given by

$$p(k|Y(t)) = \frac{\omega_k p(Y(t)|\boldsymbol{\mu}_{Y,k}, \boldsymbol{\Sigma}_{Y,k})}{\sum_{k=1}^{K} \omega_k p(Y(t)|\boldsymbol{\mu}_{Y,k}, \boldsymbol{\Sigma}_{Y,k})}. \tag{8}$$

In a similar manner, the PRM estimate of the clean speech $\widetilde{\boldsymbol{\mu}}_X(t)$ is obtained using the same posterior probability and the corresponding clean speech mean parameter $\boldsymbol{\mu}_{X,k}$ as follows:

$$\widetilde{\boldsymbol{\mu}}_X(t) = \sum_{k=1}^{K} p(k|Y(t))\boldsymbol{\mu}_{X,k}. \tag{9}$$

Here, the mean vector of the clean speech $\boldsymbol{\mu}_{X,k}$ also corresponds to the mean vector of the noise-corrupted speech $\boldsymbol{\mu}_{Y,k}$ for the same Gaussian component index k, since $\boldsymbol{\mu}_{Y,k}$ is generated from $\boldsymbol{\mu}_{X,k}$ through the model combination as presented in Step 1. Therefore, to use the posterior probability $p(k|Y(t))$ for estimating the PRM estimate of the clean speech in this study is an acceptable procedure. Fig. 1 illustrates the proposed PRM estimation procedure presented through Step 1 and 2. Here, it is assumed that the feature vector consists of two components (i.e., 2-D feature vector) and the GMM for the speech model is modeled as three Gaussian components.

C. Step 3: Mask Estimation

In this final step, the mask of the mth frequency band at time t is determined by assessing the difference of the mth PRM components of the noise-corrupted and speech in the log-spectral domain as follows:

$$\widetilde{\mu}_Y^{\{ls\}}(t,m) - \widetilde{\mu}_X^{\{ls\}}(t,m) \underset{\text{reliable}}{\overset{\text{unreliable}}{\gtrless}} \zeta_{\text{PRM}} \tag{10}$$

where

$$\widetilde{\boldsymbol{\mu}}_X^{\{ls\}}(t) = \mathbf{C}^{-1}\widetilde{\boldsymbol{\mu}}_X(t),\ \widetilde{\boldsymbol{\mu}}_Y^{\{ls\}}(t) = \mathbf{C}^{-1}\widetilde{\boldsymbol{\mu}}_Y(t). \tag{11}$$

The threshold for the PRM-based mask estimation ζ_{PRM} is empirically found in a similar manner as that seen in the spectral subtraction based method.

As presented, the proposed PRM-based mask estimation utilizes a difference of the PRM estimates of the noise-corrupted and clean speech which are obtained using posterior probabilities of the input speech $Y(t)$. We believe that to employ these PRM estimates for mask estimation will be more reliable, compared to conventional mask estimation which relies on noise and speech estimates via spectral subtraction. As seen in (9), the PRM estimate of the clean speech is estimated by a weighted sum of the mean parameters of the clean speech which are obtained through clean speech training as represented by (3). Therefore, this procedure will reduce the risk of over or under

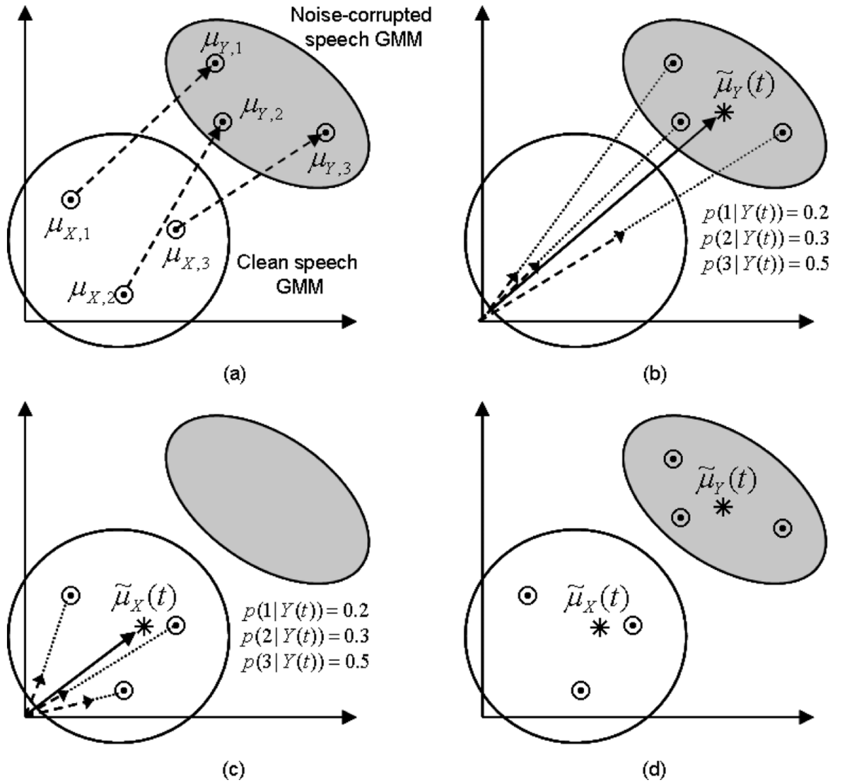

Fig. 1. Illustration of the procedure of obtaining the posterior-based representative mean estimates for input noisy speech and clean speech. (a) Noise-corrupted speech GMM generation by model combination. (b) PRM estimation for $\tilde{\mu}_Y(t)$. (c) PRM estimation for $\tilde{\mu}_X(t)$. (d) Final PRM estimates.

subtraction which mostly originates from incorrect estimation of the noise spectrum and results in degraded performance in the conventional spectral subtraction method.

Fig. 2 shows plots of (a) input noise-corrupted speech and estimated clean speech in the log-spectral domain which are used for the spectral subtraction-based method, and (b) PRM estimates of input speech and clean speech for the proposed PRM-based method. In the plots of (a), the estimates of clean speech (plain solid line) are considerably smaller compared to the original clean speech components (dashed line) for the Mel-filterbank index 10, 11, and 14 to 23. These are results of over-subtraction or taking a floor factor due to incorrect estimation of the background noise signal. In particular, the frequency bands of index 10, 11, and 14 to 17 should be determined as reliable components, since the noise corrupted speech components are still very similar with the original clean speech components in the log-spectral level. However, in this example, the small values of the estimated clean speech lead to decision of unreliable components, which represent incorrect mask information. We can see the PRM estimates for clean speech (plain line) in (b) are closer to the PRM estimates of noise-corrupted speech (plus line) in the log-spectral level for the index 10, 11, and 14 to 17, which will result in more accurate mask estimation.

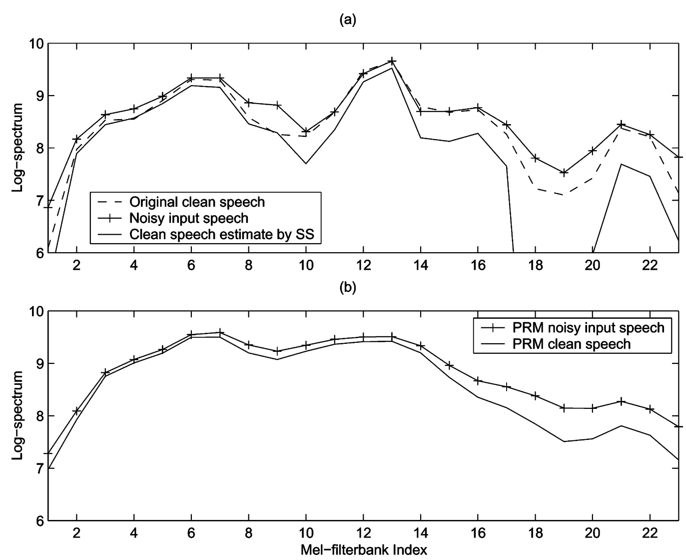

Fig. 2. Example of spectral estimates in log-spectral domain for mask estimation. (a) Input noise-corrupted speech and estimated clean speech for the spectral subtraction-based method. (b) PRM estimates of input speech and clean speech for the proposed PRM-based method.

The PRM estimate of the noise-corrupted speech is obtained using the noise-corrupted speech mean parameters which are estimated by the model combination process as presented in Step 1. As a consequence, the obtained model parameters of the noise-corrupted speech model by model combination should reflect the variance of the noise signal $\mathbf{\Sigma}_N$. Although the noise model $(\boldsymbol{\mu}_N, \mathbf{\Sigma}_N)$ is estimated from non-speech segments in this study, the obtained noise variance would represent the range of change in the noise signal during speech to some extent. Therefore, compared to conventional spectral subtraction based method which reflects a "static" noise estimate, the PRM estimates of noise-corrupted and clean speech employed in this proposed method are expected to be a more reliable representation for the noise corruption process with background noise signals which change in characteristics during the input speech utterance duration. Furthermore, in the proposed method, estimation of the models for clean speech, noise, and noisy speech as well as the posterior probability are conducted in the cepstral domain. The cepstral coefficients are less correlated with each other than the log-spectral coefficients, leading to more accurate model estimation for small data sizes with diagonal covariance matrix.

IV. Missing-Feature Reconstruction[2]

A cluster-based missing-feature reconstruction method was previously proposed by Raj, *et al.* [25]. The method restores unreliable spectral parts of input speech using known distributions of clean speech and reliable regions determined by masks. The distribution of the log-spectra of clean speech X is modeled by a Gaussian mixture with K clusters

$$p(X) = \sum_{k=1}^{K} \omega_k \mathcal{N}(X; \boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k}). \tag{12}$$

Suppose that a clean speech vector $X(t)$ has reliable components $X_r(t)$ with the latent original components in an unreliable (*i.e., missing*) region $X_u(t)$. That is, $X(t) = [X_r(t) X_u(t)]$. The reliable component $X_r(t)$ is identical to the corresponding observation $Y_r(t)$. The cluster k of the clean speech model is determined by the posterior probability. Since $X(t)$ contains unreliable elements, the marginal computation is applied by integrating out their dependency:

$$\widehat{k} = \arg\max_{k} \left\{ P(k) \int_{-\infty}^{Y_u(t)} P(X(t) \mid k) dX_u(t) \right\} \tag{13}$$

where $Y_u(t)$ represents the observed value of the unreliable parts and is assumed to be greater than $X_u(t)$ because it is corrupted by additive background noise. Finally, the unreliable part $X_u(t)$ is reconstructed using bounded MAP estimation based on the observations in the reliable regions $X_r(t)$ with the model parameters of the cluster $\widehat{k}$ selected by (13), and an upper bound $Y_u(t)$ as follows [25]:

$$\widetilde{X}_u(t) = \arg\max_{X_u(t)} \left\{ P(X_u(t) | X_r(t), \boldsymbol{\mu}_{X,\widehat{k}}, \boldsymbol{\Sigma}_{X,\widehat{k}}, X_u(t) \le Y_u(t)) \right\}. \tag{14}$$

Fig. 3 summarizes the resulting block diagram of the missing-feature reconstruction scheme employing the PRM-based mask estimation method proposed in this study.

V. Experimental Results

Our evaluations of the proposed method are performed within the Aurora 2.0 evaluation framework which was provided by the European Language Resources Association (ELRA) [38]. The task is connected English-language digits consisting of eleven words, with each whole word represented by a continuous-density HMM with 16 states and three mixtures per state. The feature extraction algorithm suggested by the European Telecommunication Standards Institute (ETSI) is employed for all experiments [39]. An analysis window of 25-ms duration is used with a 10-ms skip rate for 8-kHz speech data. The computed 23 Mel-filterbank outputs are transformed to 13 cepstrum coefficients including c0 (i.e., c0-c12). The first- and second-order time derivatives are also included, so the feature vector is 39-dimensional.

[2]In this section, all feature vectors and model parameters for missing-feature reconstruction are represented in the log-spectral domain. The symbol $\{ls\}$ has been omitted here.

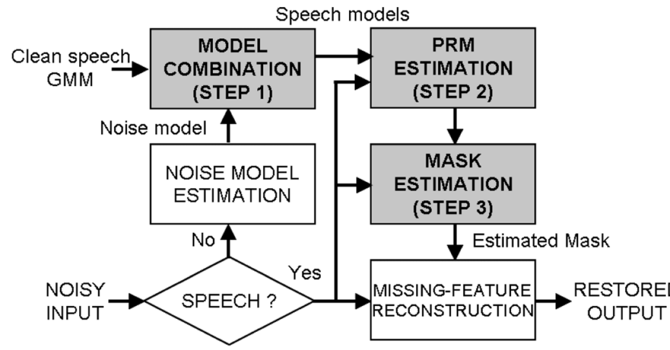

Fig. 3. Block diagram of the missing-feature reconstruction scheme employing the proposed PRM-based mask estimation method. The data flows for mask estimation and missing-feature reconstruction are in the cepstral domain and the log-spectral domain respectively.

The HMMs of the speech recognizer were trained using a database that contains 8440 utterances of clean speech. In order to evaluate the performance under various types of background noise conditions, car noise and speech babble condition were selected from the Aurora 2.0 test database, and new test data sets were generated by adding factory noise and background music samples to clean speech samples. The factory noise sample was obtained from NOISEX92 [3], [40], and the background music samples consist of prelude parts of ten Korean popular songs with varying degrees of beat and tempo. Each test set consists of 1001 samples at five different SNRs (i.e., 0, 5, 10, 15, and 20 dB), resulting in a total of 20 kinds of background noise conditions.

A. Performance of Baseline and Conventional Methods

The performance of the baseline system (no compensation) was examined with comparison to several conventional pre-processing methods in terms of speech recognition performance. Spectral subtraction (SS) [41] combined with cepstral mean normalization (CMN) was selected as one of the conventional algorithms. This represents one of the most commonly used techniques for additive noise suppression and removal of channel distortion, respectively. We also evaluated a feature compensation method, vector Taylor series (VTS) for performance comparison, where the noisy speech GMM is adaptively estimated using the EM algorithm over each test utterance [15]. The advanced front-end (AFE) algorithm developed by ETSI was also evaluated as one of the state-of-the-art methods, which contains an iterative Wiener filter and blind equalization [42]. Table I demonstrates speech recognition performance word error rate (WER) of the baseline system and conventional algorithms on

TABLE I
RECOGNITION PERFORMANCE OF BASELINE SYSTEM AND CONVENTIONAL METHODS (WER, %)

Car Noise	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Baseline	88.07	63.91	27.71	8.38	2.92	38.20
SS+CMN	53.30	19.15	6.53	2.89	2.30	16.83
VTS	83.30	44.23	12.82	4.06	2.33	29.35
VTS+SS	45.81	17.00	5.73	2.77	2.24	14.71
AFE	18.25	7.78	3.55	2.00	1.37	6.59
Factory Noise	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Baseline	83.82	54.87	21.83	5.89	2.52	33.79
SS+CMN	43.94	18.70	6.94	2.92	2.06	14.91
VTS	72.09	35.89	11.92	3.53	2.15	25.10
VTS+SS	38.59	17.56	7.43	3.25	2.00	13.77
AFE	20.60	8.87	3.75	1.90	1.35	7.29
Speech Babble	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Baseline	88.88	71.13	44.38	21.13	7.47	46.60
SS+CMN	56.53	27.21	10.91	4.78	2.66	20.42
VTS	72.04	37.82	12.58	3.93	1.90	25.65
VTS+SS	55.83	25.51	9.49	4.05	2.57	19.49
AFE	42.17	19.41	8.13	3.99	1.87	15.11
Background Music	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Baseline	74.27	51.34	28.11	12.19	4.84	34.15
SS+CMN	56.59	32.77	16.26	8.64	4.35	23.72
VTS	58.28	31.04	13.82	6.17	3.12	22.49
VTS+SS	54.67	31.60	16.08	8.92	4.69	23.19
AFE	44.43	25.55	11.72	6.76	2.99	18.29

the four types of background noise conditions with different SNRs. From the results, it can be seen that the AFE algorithm showed the best performance among the considered conventional methods and VTS combined with SS also showed better performance compared to SS + CMN.

B. *Posterior-Based Representative Mean-Based Mask Estimation*

Here, we present analysis of performance of the proposed posterior-based representative mean based mask estimation method. For the PRM-based mask estimation method in the all experiments of this paper, a 128-mixture GMM and a single Gaussian pdf for speech and noise models, respectively, were used both with diagonal covariance. Fig. 4 shows distributions of the difference values which are used for comparison to the threshold in the mask estimation methods, that are the terms of the left-hand side of (1) and (10), respectively. The difference value for the spectral subtraction based method is a subtraction of the estimated clean speech from the input noisy speech in the log-spectral domain (i.e., $y^{\{ls\}}(t,m) - \widetilde{x}^{\{ls\}}(t,m)$). The value for the PRM-based method is a subtraction of the PRM estimate of clean speech from the PRM estimate of the noise corrupted speech (i.e., $\widetilde{\mu}_Y^{\{ls\}}(t,m) - \widetilde{\mu}_X^{\{ls\}}(t,m)$). The plots in Fig. 4 were generated using the car noise condition at 5-dB SNR. The solid circle and empty circle represent mean values (i.e., average) of the difference values at each Mel-filterbank index for reliable components and unreliable components, respectively, also showing their standard deviations with small bars. The thresholds for mask decision (i.e., ζ_{SS} and ζ_{PRM}) could be formulated between the mean values for reliable and unreliable components. From these plots, it can be seen that the distributions of the difference values of the proposed PRM-based method formulate more distinctively for reliable and unreliable components, compared to the spectral subtraction based method. We believe that this more separable property of the proposed PRM-based mask estimation method will result in improved performance compared to the SS-based method.

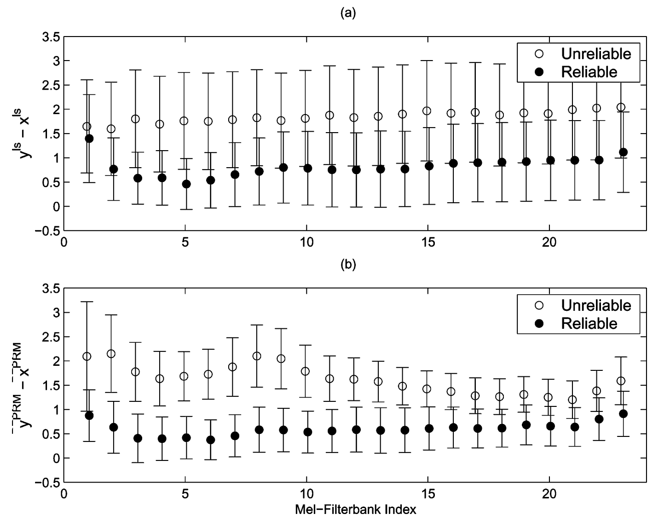

Fig. 4. Distributions of the left-hand side terms of equations (1) and (10) in car noise condition at 5-dB SNR: solid circles and empty circles indicate mean values of the distributions for reliable and unreliable components respectively. (a) SS-based mask estimation. (b) PRM-based mask estimation.

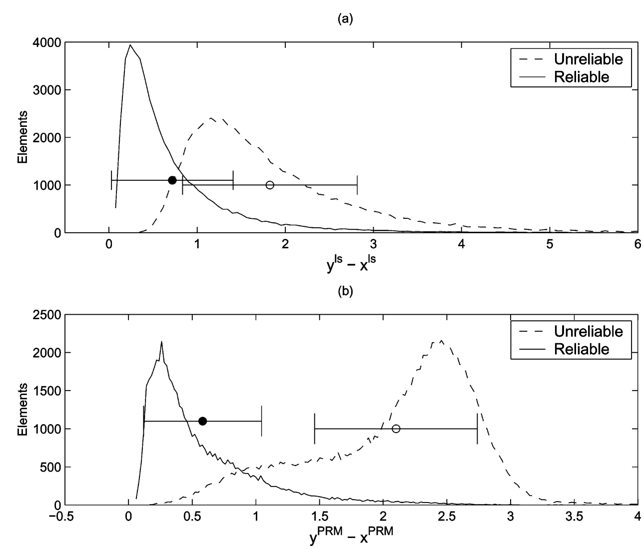

Fig. 5. Histograms of the difference values of the 8th Mel-filterbank index in car noise condition at 5-dB SNR. (a) SS-based mask estimation. (b) PRM-based mask estimation.

Fig. 5 displays a detailed illustration of the distributions (i.e., histograms) of the difference values at the 8th Mel-filterbank index for the SS-based and PRM-based mask estimation methods. Here, the mean values and their standard deviations were presented also, which are matched to ones presented at the 8th index in Fig. 4. We can also see the distributions of the difference values for reliable and unreliable components are more separable in (b) the proposed PRM-based method compared to (a) the SS-based method. Fig. 6 shows a comparison of the distributions of the difference values for the SS-based and PRM-based methods for speech babble noise conditions. From the comparison of the plots we also can see the proposed PRM-based method represents more separable distributions.

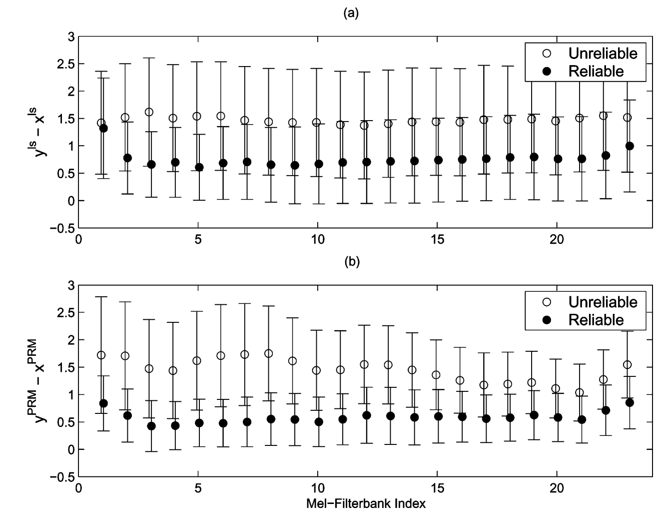

Fig. 6. Distributions of the left-hand-side terms of the equations (1) and (10) in speech babble condition at 5-dB SNR: solid circles and empty circles indicate mean values of the distributions for reliable and unreliable components, respectively. (a) SS-based mask estimation. (b) PRM-based mask estimation.

TABLE II
RECOGNITION PERFORMANCE OF MISSING-FEATURE RECONSTRUCTION (MF) EMPLOYING SS-BASED (SSM) AND PRM-BASED (PRM) MASK ESTIMATION METHODS IN FOUR TYPES OF BACKGROUND NOISE CONDITIONS (WER, %)

Car	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Oracle-MF	16.13	7.61	4.12	3.40	2.71	6.79
SSM-MF	39.25	21.59	8.80	3.79	2.74	15.23
PRM-MF	**35.97**	**13.93**	**5.64**	**3.55**	**2.77**	**12.37**
(Relative)	**(+8.36)**	**(+35.48)**	**(+35.91)**	**(+6.33)**	**(-1.09)**	**(+17.00)**
Factory	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Oracle-MF	13.33	7.34	3.75	2.61	2.18	5.84
SSM-MF	44.70	24.72	10.96	3.96	2.95	17.46
PRM-MF	**36.44**	**17.01**	**6.32**	**3.22**	**2.12**	**13.02**
(Relative)	**(+18.48)**	**(+31.19)**	**(+42.34)**	**(+18.69)**	**(+28.14)**	**(+27.77)**
Babble	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Oracle-MF	14.00	8.37	4.87	3.63	2.78	6.73
SSM-MF	60.31	33.37	15.02	7.07	3.84	23.92
PRM-MF	**57.74**	**27.03**	**9.16**	**4.32**	**2.75**	**20.20**
(Relative)	**(+4.26)**	**(+19.00)**	**(+39.01)**	**(+38.90)**	**(+28.39)**	**(+25.91)**
Music	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Oracle-MF	10.43	5.43	3.52	2.25	1.94	4.71
SSM-MF	64.33	39.59	19.53	8.73	4.44	27.32
PRM-MF	**41.65**	**23.57**	**10.83**	**5.25**	**2.25**	**16.71**
(Relative)	**(+35.26)**	**(+40.46)**	**(+44.55)**	**(+39.86)**	**(+49.32)**	**(+41.89)**

C. Missing-Feature Speech Recognition Performance Employing Mask Estimation

Tables II and III show recognition performance using the missing-feature (MF) reconstruction method employing the mask estimation methods for four types of background noise conditions. In missing-feature reconstruction for all our experiments, a 23rd-order log-spectral coefficients (i.e., log of Mel-filterbank output) were used for the feature vector and a 64-mixture GMM with a full covariance was employed. The reconstructed feature in the log-spectral domain is transformed to the cepstral coefficients and then submitted to the speech recognizer with a clean condition trained HMM. The WERs with the "Oracle" mask (i.e., Oracle-MF) represent the recognition performance with perfect knowledge of the reliable/unreliable regions, providing an upper bound on performance for evaluating mask estimation methods. The Oracle mask was generated by comparing the noise-corrupted speech signal to the original clean speech at the log-spectrum level. For noise estimates for both the spectral subtraction based and the proposed PRM-based mask estimation methods, we used the silence (i.e., non-speech) duration at the beginning and end parts of each utterance which consists of a total of 24 frames.

TABLE III
RECOGNITION PERFORMANCE OF MISSING-FEATURE RECONSTRUCTION (MF) COMBINED WITH SPECTRAL SUBTRACTION (SS), EMPLOYING SS-BASED (SSM) AND PRM-BASED (PRM) MASK ESTIMATION METHODS IN FOUR TYPES OF BACKGROUND NOISE CONDITIONS (WER, %)

Car	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
SSM-MF+SS	38.14	15.30	6.38	3.61	2.74	13.23
PRM-MF+SS	**25.08**	**8.86**	**4.62**	**3.40**	**2.62**	**8.92**
(Relative)	**(+34.24)**	**(+42.09)**	**(+27.59)**	**(+5.82)**	**(+4.38)**	**(+22.82)**
Factory	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
SSM-MF+SS	44.95	20.23	7.52	2.95	2.14	15.57
PRM-MF+SS	**22.23**	**10.35**	**5.07**	**3.19**	**2.33**	**8.63**
(Relative)	**(+50.55)**	**(+48.84)**	**(+32.58)**	**(-8.14)**	**(-6.88)**	**(+23.39)**
Babble	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
SSM-MF+SS	58.10	29.47	12.76	6.47	4.23	22.21
PRM-MF+SS	**42.78**	**17.08**	**6.80**	**3.99**	**2.90**	**14.71**
(Relative)	**(+26.37)**	**(+42.04)**	**(+46.71)**	**(+38.33)**	**(+31.44)**	**(+36.98)**
Music	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
SSM-MF+SS	58.72	34.43	17.53	8.92	4.10	24.74
PRM-MF+SS	**37.92**	**22.68**	**10.95**	**5.89**	**3.33**	**16.15**
(Relative)	**(+35.42)**	**(+34.13)**	**(+37.54)**	**(+33.97)**	**(+18.78)**	**(+31.97)**

TABLE IV
THRESHOLD VALUES USED FOR THE MASK ESTIMATION IN THE EXPERIMENTS

	Car	Factory	Babble	Music
SSM-MF (ζ_{SS})	0.78	0.74	0.52	0.75
SSM-MF+SS (ζ_{SS})	1.12	1.78	0.49	0.73
PRM-MF{+SS} (ζ_{PRM})	0.8 ($f \leq 1$ kHz), 1.2 ($f > 1$ kHz)			

As shown in these results, there were significant relative improvements in WER by employing the proposed PRM-based mask estimation method (PRM-MF). We obtained $+17.00\%$, $+27.77\%$, $+25.91\%$, and $+41.89\%$ average relative improvements[3] in WER for car, factory, babble, and music noise conditions respectively, compared to the reconstruction method with the spectral subtraction-based mask estimation (SSM-MF). The threshold values (i.e., ζ_{SS} and ζ_{PRM}) for both SS-based and PRM-based mask estimation were determined in a empirical way to achieve the best performance in an average WER for all SNRs. It was found that the PRM-based method shows consistently improved performance with a single frequency-dependent threshold for all noise conditions. We used 0.8 for the frequency range below 1 kHz and 1.2 for the range over 1 kHz of input speech signal. For the spectral subtraction based method, a similar frequency-dependent threshold did not show a consistent performance improvement. The used threshold values are presented in Table IV.

We found that the missing-feature reconstruction method produces a significant improvement in WER when combined with the conventional spectral subtraction prior to the missing-feature processing. For the combination scheme, input speech signal is first enhanced using the spectral subtraction, and then fed to the missing-feature reconstruction method. The mask estimation methods are also applied to the enhanced input speech signal. Table III shows the performance of missing-feature reconstruction combined with spectral subtraction employing the two types of mask estimation methods.

[3]The average relative improvement is computed by taking the average of the obtained relative improvements.

TABLE V
RECOGNITION PERFORMANCE OF MISSING-FEATURE RECONSTRUCTION (MF) WITH PRM-BASED (PRM) MASK ESTIMATION METHODS EMPLOYING VOICE ACTIVITY DETECTOR (VAD) IN FOUR TYPES OF BACKGROUND NOISE CONDITIONS (WER, %); RELATIVE IMPROVEMENTS ARE OBTAINED BY COMPARISON TO SSM-MF OF TABLE II AND SSM-MF + SS OF TABLE III

Car	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
V+PRM-MF	34.80	14.49	6.23	3.52	3.13	12.43
(Relative)	(+11.34)	(+32.89)	(+29.20)	(+7.12)	(-14.23)	(+13.26)
V+PRM-MF+SS	23.62	9.39	5.13	3.19	3.37	8.94
(Relative)	(+38.07)	(+38.63)	(+19.59)	(+11.63)	(-22.99)	(+16.99)
Factory	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
V+PRM-MF	36.05	16.79	7.61	3.65	2.64	13.35
(Relative)	(+19.35)	(+32.08)	(+30.57)	(+7.83)	(+10.51)	(+20.07)
V+PRM-MF+SS	25.48	11.82	6.23	3.62	2.89	10.01
(Relative)	(+43.31)	(+41.57)	(+17.15)	(-22.71)	(-32.57)	(+9.35)
Babble	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
V+PRM-MF	56.47	24.97	8.04	4.72	3.69	19.58
(Relative)	(+6.37)	(+25.17)	(+46.47)	(+33.24)	(+3.91)	(+23.03)
V+PRM-MF+SS	43.80	16.90	6.35	4.05	3.54	14.93
(Relative)	(+24.61)	(+42.65)	(+50.24)	(+37.40)	(+16.31)	(+34.24)
Music	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
V+PRM-MF	42.05	21.54	11.82	5.74	3.52	16.93
(Relative)	(+34.63)	(+45.59)	(+39.48)	(+34.25)	(+20.72)	(+34.93)
V+PRM-MF+SS	36.66	19.44	10.77	5.37	3.86	15.22
(Relative)	(+37.57)	(+43.54)	(+38.56)	(+39.80)	(+5.85)	(+33.06)

As shown in these results, there were also consistently significant relative improvements in WER by employing the proposed PRM-based mask estimation method. We obtained $+22.82\%$, $+23.39\%$, $+36.98\%$, and $+31.97\%$ average relative improvements in WER for the car, factory, babble, and music conditions, respectively, compared to $\text{SSM-MF} + \text{SS}$.

For real-life application, we employed a voice activity detection (VAD) algorithm for noise estimation for the PRM-based mask estimation, which does not require prior knowledge of non-speech segments locations of input speech. Here we employed a simple VAD method which is based on quantile statistics of energy values. In our method, the energy values of all frames of every input utterance are sorted and then a median value is selected for the threshold to decide speech or non-speech frames. The recognition performance of the missing-feature reconstruction with the proposed PRM-based mask estimation method employing the VAD algorithm $(\text{V} + \text{PRM-MF}\{+\text{SS}\})$ is presented in Table V with relative improvements obtained by comparing to the $\text{SSM-MF}\{+\text{SS}\}$ of Tables II and III. These results prove that to employ the VAD algorithm for the PRM-based method still results in considerable improvements compared to $\text{SSM-MF}\{+\text{SS}\}$, although there was performance degradation at higher SNRs for car and factory noise conditions. We believe that more reliable VAD algorithm will compensate the performance degradation and bring more improved performance.

The comparison of performance of the missing-feature method employing the proposed PRM-based mask estimation with other conventional methods are summarized for different types of background noise conditions (Table VI) and different SNR conditions (Table VII). From these results, we can see that the proposed PRM-based method showed $+28.14\%$ and $+28.79\%$ average relative improvements for all noise conditions compared to the SS-based method, solely used and combined with with spectral subtraction, respectively. We note that the average WERs of the missing-feature method employing the proposed PRM-based mask estimation method outperforms the AFE[4] for babble (14.71% versus 15.11%) and background music (16.15% versus 18.29%) conditions. By employing the VAD algorithm, $+22.82\%$ and $+23.41\%$ average relative improvements compared to the $\text{SSM-MF}\{+\text{SS}\}$ were obtained for all noise conditions. It also provides more effective performance for babble (14.93% versus 15.11%) and background music (15.22% versus 18.29%) noise conditions compared to the AFE algorithm. It is worth to note that the proposed PRM-based mask estimation method shows effective performance with a single frequency-dependent threshold as shown in Table IV which is independent of the noise condition, while selection of the threshold for the conventional SS-based method is highly sensitive to the background noise type to produce the best performance.

TABLE VI
PERFORMANCE COMPARISON IN WER (%) IN FOUR TYPES OF BACKGROUND NOISE CONDITIONS AS AVERAGE OVER ALL SNRS; 0, 5, 10, 15 AND 20 dB

	Car	Factory	Babble	Music	Avg.
Baseline	38.20	33.79	46.60	34.15	38.18
SS+CMN	16.83	14.91	20.42	23.72	18.97
VTS+SS	14.71	13.77	19.49	23.19	17.79
AFE	6.59	7.29	15.11	18.29	11.82
SSM-MF	15.23	17.46	23.92	27.32	20.98
PRM-MF	**12.37**	**13.02**	**20.20**	**16.71**	**15.58**
(Relative)	**(+17.00)**	**(+27.77)**	**(+25.91)**	**(+41.89)**	**(+28.14)**
V+PRM-MF	**12.43**	**13.35**	**19.58**	**16.93**	**15.57**
(Relative)	**(+13.26)**	**(+20.07)**	**(+23.03)**	**(+34.93)**	**(+22.82)**
SSM-MF+SS	13.23	15.57	22.21	24.74	18.94
PRM-MF+SS	**8.92**	**8.63**	**14.71**	**16.15**	**12.10**
(Relative)	**(+22.82)**	**(+23.39)**	**(+36.98)**	**(+31.97)**	**(+28.79)**
V+PRM-MF+SS	**8.94**	**10.01**	**14.93**	**15.22**	**12.27**
(Relative)	**(+16.99)**	**(+9.35)**	**(+34.24)**	**(+33.06)**	**(+23.41)**

TABLE VII
PERFORMANCE COMPARISON IN WER (%) IN DIFFERENT SNR CONDITIONS AS AVERAGE OVER ALL FOUR BACKGROUND NOISE TYPES

	0 dB	5 dB	10 dB	15 dB	20 dB	Avg.
Baseline	83.76	60.31	30.51	11.90	4.44	38.18
SS+CMN	52.59	24.46	10.16	4.81	2.84	18.97
VTS+SS	48.73	22.92	9.68	4.75	2.88	17.79
AFE	31.36	15.40	6.79	3.66	1.90	11.82
SSM-MF	52.15	29.82	13.58	5.89	3.49	20.98
PRM-MF	**42.95**	**20.39**	**7.99**	**4.08**	**2.47**	**15.58**
(Relative)	**(+16.59)**	**(+31.53)**	**(+40.45)**	**(+25.94)**	**(+26.19)**	**(+28.14)**
V+PRM-MF	**42.34**	**19.45**	**8.43**	**4.41**	**3.24**	**15.58**
(Relative)	**(+17.92)**	**(+33.93)**	**(+36.43)**	**(+20.61)**	**(+5.23)**	**(+22.82)**
SSM-MF+SS	49.98	24.86	11.05	5.49	3.31	18.94
PRM-MF+SS	**32.00**	**14.74**	**6.86**	**4.12**	**2.80**	**12.10**
(Relative)	**(+36.64)**	**(+41.77)**	**(+36.10)**	**(+17.50)**	**(+11.93)**	**(+28.79)**
V+PRM-MF+SS	**32.39**	**14.39**	**7.12**	**4.06**	**3.42**	**12.27**
(Relative)	**(+35.89)**	**(+41.60)**	**(+31.39)**	**(+16.53)**	**(-8.35)**	**(+23.41)**

D. Real-Life In-Vehicle Condition: CU-Move Corpus

The proposed mask estimation method for missing-feature reconstruction was also evaluated on a real-life in-vehicle conditions obtained from the CU-Move corpus [7]. The CU-Move project was designed to develop reliable car navigation systems employing a mixed-initiative dialog. This requires robust speech recognition across changing acoustic conditions. The CU-Move database consists of five parts: 1) command and control words; 2) digit strings of telephone and credit numbers; 3) street names and addresses; 4) phonetically-balanced sentences, and 5) Wizard of Oz interactive navigation conversations. A total of 500 speakers, balanced across gender and age,

[4]AFE showed the best performance when used in isolation without SS.

TABLE VIII
RECOGNITION PERFORMANCE IN WER (%) COMPARISON FOR THE CU-MOVE CORPUS: RELATIVE IMPROVEMENT COMPARED TO SSM-MF IS SHOWN IN A PARENTHESIS

Baseline	70.02	
SS+CMN	39.90	
VTS+SS	30.98	
AFE	31.45	
SSM-MF+SS	34.18	
PRM-MF+SS	**29.79**	**(+12.84)**
V+PRM-MF+SS	**30.95**	**(+9.45)**

produced over 600 GB of data during a six-month collection effort across the United States. The database and noise conditions are discussed in detail in [7]. For the evaluation in this study, we selected 949 utterances (length of 1 hour and 40 min) spoken by 20 different speakers (9 males and 11 females), which were collected in Minneapolis, MN. The test samples represent an average 8.48 dB[5] SNR calculated by the NIST STNR Speech Quality Assurance software [43].

Table VIII shows the performance evaluation of the proposed PRM-based mask estimation for missing-feature reconstruction on the CU-Move corpus. These results demonstrate that missing-feature reconstruction with the proposed PRM-based mask estimation brings consistent improvement compared to the SSM-MF on the real-life in-vehicle condition as well, resulting in $+12.84\%$ relative improvement when combined with spectral subtraction. By employing the identical VAD algorithm presented in Section V-C, we obtained $+9.45\%$ of relative improvement. Here also the same threshold value as presented in Table III for the PRM-based method was employed. These results also show that the performance of the proposed $\mathrm{PRM\text{-}MF} + \mathrm{SS}$ schemes slightly outperform both $\mathrm{VTS} + \mathrm{SS}$ and AFE. The results here prove that the proposed PRM-based mask estimation method could be applicable to real-life in-vehicle conditions to improve performance of speech recognition.

VI. CONCLUSION

This study has proposed a novel mask estimation method for missing-feature reconstruction to improve speech recognition in various types of background noise conditions. In the proposed method, a posterior-based representative mean estimate was utilized to determine the reliability of the input speech spectrum, which is obtained as a weighted sum of mean parameters of the speech model using the posterior probability. To obtain the noise-corrupted speech model, a model combination method was employed, which was previously proposed for feature compensation in our past study. Experimental results demonstrated that the proposed mask estimation method provides more separable distributions for the reliable/unreliable component classifier compared to the conventional mask estimation method. The recognition performance was evaluated using the Aurora 2.0 framework over four types of background noise conditions (e.g., car, factory, speech babble and background music) and the CU-Move corpus which is for real-life in-vehicle conditions. The performance evaluation showed that the proposed mask estimation method is considerably more effective at increasing speech recognition performance in various types of background noise conditions, compared to the conventional mask estimation method which is based on spectral subtraction. By employing the proposed PRM-based mask estimation for missing-feature reconstruction, we obtained $+23.41\%$ and $+9.45\%$ average relative improvements in WER for all four types of noise conditions and the CU-Move corpus, respectively, compared to conventional mask estimation methods. It is noted that the proposed missing-feature method with spectral subtraction outperformed the ETSI AFE algorithm in speech babble and background music for the Aurora 2.0 framework and CU-Move corpus. These advancements contribute to the increased viability of missing-feature theory for robust speech systems in time-varying noisy environments.

[5]0 dB and 5 dB SNR test samples of the car noise condition of the Aurora 2.0 corpus show 7.15 dB and 11.66 dB average SNRs, respectively, using the NIST STNR tool.

REFERENCES

[1] W. Kim and J. H. L. Hansen, "Mask estimation employing posterior-based representative mean for missing-feature speech recognition with time-varying background noise," in *Proc. IEEE ASRU'09*, Merano, Italy, Dec. 2009, pp. 194–198.
[2] W. Kim and J. H. L. Hansen, "Feature compensation in the cepstral domain employing model combination," *Speech Commun.*, vol. 51, no. 2, pp. 83–96, 2009.
[3] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," in *Tech. Rep., Speech Res. Unit, Defense Res. Agency*, Malvern, U.K., 1992, (Available from NOISEX-92 CD-ROMS).
[4] H. Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The Speechdat-Car multilingual speech databases for in-car applications: Some first validation results," in *Proc. Eurospeech'99*, Sep. 1999.
[5] T. Crystal, A. Schmidt-Nelson, and E. Marsh, "Speech in noisy environments (SPINE) adds news dimension to speech recognition R&D," in *Proc. HLT Conf.*, San Diego, CA, Mar. 2002.
[6] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. H. L. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *Proc. IEEE Intell. Veh. Conf.*, 2007, pp. 566–569.
[7] J. H. L. Hansen, X. Zhang, M. Akbacak, U. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul, "CU-move: Advances for in-vehicle speech systems for route navigation," in *DSP for In-Vehicle and Mobile Systems*, Abut, J. H. L. Hansen, and Takeda, Eds. New York: Springer., 2004, ch. 2.
[8] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller Jr., A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a National Gallery of the Spoken Word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, Sep. 2005.
[9] W. Kim and J. H. L. Hansen, "Speechfind for CDP: Advances in spoken document retrieval for the U.S. collaborative digitization program," in *Proc. IEEE ASRU2007*, 2007, pp. 687–692.
[10] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, no. 2, pp. 151–170, 1996.
[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[12] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[13] J. H. L. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[14] J. H. L. Hansen, "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 598–614, Oct. 1994.
[15] P. J. Moreno, B. Raj, and R. M. Stern, "Data-driven environmental compensation for speech recognition: A unified approach," *Speech Commun.*, vol. 24, no. 4, pp. 267–285, 1998.
[16] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Commun.*, vol. 37, pp. 231–248, 2002.

[17] V. Stouten, H. Van hamme, and P. Wambacq, "Joint removal of additive and convolutional noise with model-based feature enhancement," in *Proc. ICASSP'04*, 2004, pp. 949–952.
[18] A. Sasou, T. Tanaka, S. Nakamura, and F. Asano, "HMM-based feature compensation methods: An evaluation using the Aurora2," in *Proc. ICSLP'04*, 2004, pp. 121–124.
[19] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.
[21] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.
[22] J. Barker, M. Cooke, and P. Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. Eurospeech'01*, 2001, pp. 213–216.
[23] M. Cook, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, 2001.
[24] K. J. Palomaki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with missing data automatic speech recognition," *Speech Commun.*, vol. 43, pp. 123–142, 2004.
[25] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, 2004.
[26] H. Van Hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP'04*, May 2004, pp. 213–216.
[27] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101–116, Sep. 2005.
[28] W. Kim and J. H. L. Hansen, "Missing-feature reconstruction for band-limited speech recognition in spoken document retrieval," in *Proc. Interspeech'06*, Sep. 2006, pp. 2306–2309.
[29] W. Kim and J. H. L. Hansen, "Time-frequency correlation based missing-feature reconstruction for robust speech recognition in band-restricted conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1292–1304, Sep. 2009.
[30] A. Vizinho, P. Green, M. M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-Noise estimation for robust ASR: An integrated study," in *Proc. Eurospeech'99*, Sep. 1999, pp. 2407–2410.
[31] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing-feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, 2004.
[32] W. Kim, R. M. Stern, and H. Ko, "Environment-independent mask estimation for missing-feature reconstruction," in *Proc. Interspeech'05*, Sep. 2005, pp. 2637–2640.
[33] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. ICASSP'06*, May 2006, pp. 305–308.
[34] P. Jancovic, M. Kokuer, and F. Murtagh, "High-likelihood model based on reliability statistics for robust combination of features: Application to noisy speech recognition," in *Proc. Eurospeech'03*, 2003, pp. 2161–2164.
[35] S. Harding, J. Barker, and G. J. Brown, "Mask estimation based on sound localisation for missing data speech recognition," in *Proc. ICASSP'05*, 2005, pp. 537–540.
[36] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, 2006.
[37] H. Park and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings," *Speech Commun.*, vol. 51, no. 1, pp. 15–25, 2009.
[38] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR2000*, 2000.
[39] *ETSI Standard Document*, ETSI ES 201 108 v1.1.2 (2000-04), 2000.
[40] [Online]. Available: http://spib.rice.edu/spib/select_noise.html
[41] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO-94*, 1994, pp. 1182–1185.
[42] *ETSI Standard Document*, ETSI ES 202 050 v1.1.1 (2002-10), 2002.
[43] NIST SPeech Quality Assurance (SPQA) Package Version 2.3 [Online]. Available: http://www.nist.gov/speech

Wooil Kim (M'06) received the B.S., M.S., and Ph.D. degrees in electronics engineering from Korea University, Seoul, in 1996, 1998, and 2003, respectively.

He has been a Research Assistant Professor in the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, since September 2007. He is also a member of the Center for Robust Speech Systems (CRSS) at UTD. Previously, he was a Research Associate at UTD (August 2005–August 2007) and a Post-Doctoral Researcher in the Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA (August 2004–August 2005) and Korea University (September 2003–August 2004). His research interests are robust speech recognition in adverse environments, acoustic modeling for large-vocabulary continuous speech recognition, and spoken document retrieval.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, GA, in 1983 and 1988, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado at Boulder (1998–2005), where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 51 (22 Ph.D., 29 M.S./M.A.) thesis candidates. He is author/coauthor of 380 journal and conference papers and eight textbooks in the field of speech processing and language technology, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), co-editor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was named IEEE Fellow in 2007 for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2005–2008; 2010–2013; elected Chair-elect in 2010), and Educational Technical Committee (2005–2008; 2008–2010). Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the IEEE Signal Processing Magazine (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council. In 2010, he was recognized as an ISCA Fellow, for contributions on "research for speech processing of signals under adverse conditions." He was recipient of The 2005 University of Colorado Teacher Recognition Award as voted on by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and served as Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.