Missing-Feature Reconstruction by Leveraging Temporal Spectral Correlation for Robust Speech Recognition in Background Noise Conditions

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

***Abstract*—This paper proposes a novel missing-feature reconstruction method to improve speech recognition in background noise environments. The existing missing-feature reconstruction method utilizes log-spectral correlation across frequency bands. In this paper, we propose to employ a temporal spectral feature analysis to improve the missing-feature reconstruction performance by leveraging temporal correlation across neighboring frames. In a similar manner with the conventional method, a Gaussian mixture model is obtained by training over the obtained temporal spectral feature set. The final estimates for missing-feature reconstruction are obtained by a selective combination of the original frequency correlation based method and the proposed temporal correlation-based method. Performance of the proposed method is evaluated on the TIMIT speech corpus using various types of background noise conditions and the CU-Move in-vehicle speech corpus. Experimental results demonstrate that the proposed method is more effective at increasing speech recognition performance in adverse conditions. By employing the proposed temporal-frequency based reconstruction method, a $+17.71\%$ average relative improvement in word error rate (WER) is obtained for white, car, speech babble, and background music conditions over 5-, 10-, and 15-dB SNR, compared to the original frequency correlation-based method. We also obtain a $+16.72\%$ relative improvement in real-life in-vehicle conditions using data from the CU-Move corpus.**

***Index Terms*—Background noise, missing-feature, robust speech recognition, temporal correlation, temporal spectral feature.**

I. Introduction

BACKGROUND noise is one of the primary factors resulting in acoustic mismatch between training and operating conditions for actual speech recognition systems, severely degrading recognition performance. Typical examples can be easily found in the corpora of UTDrive [1], CU-Move [2], the National Gallery of Spoken Word (NGSW) [3], Collaborative Digitization Program (CDP) [4], Speech Under Simulated and Actual Stress (SUSAS) including Lombard effect [5], and others, which make speech recognition technology challenging in real-life scenarios. To minimize this mismatch, extensive research have been conducted in recent decades, which include many types of speech/feature enhancement methods such as Spectral Subtraction, Cepstral Mean Normalization, and variety of feature compensation schemes [5]–[14]. 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) [15]–[17]. Recently, missing-feature methods have shown promising results [18]–[27] with some that utilize no prior knowledge of the background noise [25].

In this paper, the missing-feature method is considered as a solution to address additive background noise for speech recognition. This method depends primarily on characteristics of speech that is resistant to noise, rather than on the characteristics of the noise itself, showing its effectiveness at improving speech recognition in adverse environments [18], [19],[21]. 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 [25], [28]. The second step is to reconstruct the unreliable regions or bypass them for alternative processing.

A cluster-based reconstruction method [21] is employed as a framework for missing-feature processing of background noise corrupted speech in our study. This method restores unreliable parts of incoming speech signal using known distributions of clean speech and reliable spectral regions indicated by mask information. The existing cluster-based method [21] is designed to utilize the correlation relationship of log-spectral components across frequency bands by employing a Gaussian mixture model (GMM) with full covariance trained over the conventional log-spectral coefficients which were used as the speech feature vector. With proper knowledge of the mask, the cluster-based reconstruction method shows considerable effectiveness at increasing speech recognition performance in additive background noise conditions.

This paper represents a new effort to improve missing-feature reconstruction performance for speech recognition in background noise environments. In this paper, we leverage the available spectral correlation across neighboring frames in the missing-feature reconstruction method, by employing a temporal spectral feature analysis which is conducted using the conventional log-spectral coefficients. The missing-feature is finally reconstructed by a selective combination of the original reconstruction method and the proposed temporal correlation-based method. Prior efforts have also attempted to

Manuscript received June 30, 2009; revised December 09, 2009. Date of publication June 07, 2010; date of current version October 15, 2010. This work was supported by the USAF under a subcontract to RADC, Inc., Contract FA8750-09-C-0067 (Approved for public release. Distribution unlimited.). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

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.2010.2041698

1558-7916/$26.00 © 2010 IEEE

utilize the temporal correlation for missing-feature method, by employing a relative covariance value [21], [29] or hidden Markov model (HMM) [30]; however; such approaches are inferior to the conventional cluster-based method [29] or not suitable to our goal in this study [30]. Our recent study has proposed a time–frequency correlation based method, but it is only applicable to the band-limited speech condition [27]. Independent of the missing-feature method, many studies also have been conducted for utilizing the temporal information of speech feature to improve robustness of speech recognition [31]–[34].

This paper is organized as follows. We first review the cluster-based missing-feature reconstruction method as a framework for this study in Section II. Section III presents the proposed missing-feature reconstruction method including temporal spectral feature extraction and details of the proposed processing. Representative experimental procedures and their results are presented and discussed in Section IV. Finally, in Section V we state the main conclusions of our work.

II. Missing-Feature Reconstruction: Frequency Correlation Based Method

A cluster-based missing-feature reconstruction method was previously proposed by Raj, *et al.* [21]. The method restores unreliable spectral parts of input speech using a known distribution of clean speech and reliable regions determined by the masks. The distribution of the log-spectra of clean speech $X(t)$ is modeled by a Gaussian mixture with K clusters,

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

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

$$\hat{k} = \arg\max_{k} \left\{ P(k) \int_{-\infty}^{Y_u(t)} P(X(t) \mid k) dX_u(t) \right\} \tag{2}$$

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 $\hat{k}$ selected by (2), and an upper bound $Y_u(t)$ as follows [21]:

$$\tilde{X}_u(t) = \arg\max_{X_u(t)} \{ P(X_u(t) | X_r(t), \boldsymbol{\mu}_{X,\hat{k}}, \boldsymbol{\Sigma}_{X,\hat{k}}, X_u(t) \le Y_u(t)) \}. \tag{3}$$

(3) can be simplified into the following [29]:

$$\tilde{X}_u(t) = \boldsymbol{\mu}_{\hat{k},u} + \boldsymbol{C}_{\hat{k},ru} \cdot \boldsymbol{C}_{\hat{k},rr}^{-1} \cdot [Y_r(t) - \boldsymbol{\mu}_{\hat{k},r}] \tag{4}$$

where $\boldsymbol{C}_{\hat{k},rr}$ and $\boldsymbol{C}_{\hat{k},ru}$ are the covariance and cross-covariance matrices defined as follows:

$$\boldsymbol{C}_{\hat{k},rr} = E\{(X_r(t) - \boldsymbol{\mu}_{\hat{k},r})(X_r(t) - \boldsymbol{\mu}_{\hat{k},r})^T\} \tag{5}$$

$$\boldsymbol{C}_{\hat{k},ru} = E\{(X_r(t) - \boldsymbol{\mu}_{\hat{k},r})(X_u(t) - \boldsymbol{\mu}_{\hat{k},u})^T\} \tag{6}$$

where $\boldsymbol{\mu}_{\hat{k},r}$ and $\boldsymbol{\mu}_{\hat{k},u}$ are mean vectors of the $\hat{k}$th cluster of the reliable component $X_r(t)$ and unreliable component $X_u(t)$ of the clean speech, respectively.

In this conventional method, the unreliable parts $X_u(t)$ are reconstructed depending on the reliable components $X_r(t)$ of a frame at time t, utilizing only correlation across frequency bands (i.e., Mel-scales filterbanks). From now, we denote the reconstructed component $\tilde{X}_u(t)$ by (4) as $\tilde{X}_u^{\{f\}}(t)$ to represent *frequency correlation*, which is distinguished from the *temporal correlation* based method which is proposed in the next section.

III. Missing-Feature Recognition Leveraging Temporal Correlation

In this section, we propose a novel approach to improve performance of the conventional missing-feature reconstruction method presented in Section II. The proposed method leverages the correlation of unreliable components with reliable components from neighboring frames as well as the current frame which conventional methods utilize. Raj, *et al.* previously proposed a correlation-based reconstruction method in [21], [29]. Their proposed method employs a relative covariance value to determine a neighborhood vector which is more correlated to missing components and used for reconstruction. In their method; however, the spectrogram of the clean speech signal is considered to be a wide-sense stationary random process, so the distribution of the clean speech is estimated simply using a single Gaussian pdf. Such a simplification results in inferior performance compared to the cluster-based method which is the baseline scheme for our work in this paper [29].

Borgstrom and Alwan [30] have proposed an HMM-based estimation method of unreliable components by utilizing correlations across feature vectors and frequency channels. In their work, several noise parameters for the HMM construction are estimated from noise-corrupted speech. We believe that the performance of such an approach would be dependent on the ability of noise parameter estimation,[1] which also would rely on the types of background noise conditions. This aspect is not suitable for our purpose in this paper, where the acoustic model for missing-feature reconstruction only utilizes clean speech statistics through offline training, as suggested by the representative conventional missing-feature methods [18]–[21]. In our paper, the missing-feature reconstruction procedure is intended to be free of environment-dependent factors, once the mask information is provided. An HMM-based method is still considered to be considerably expensive in training acoustic models and computing likelihood scores compared to a GMM-based method, even though the study in [30] proposed to use a down-sampled size of HMM. In our recent study, we have proposed a new missing-feature reconstruction method for band-limited speech,

[1]It should be noted that there is no clear description on exactly how to estimate some HMM parameters in their study [30].

which utilizes the correlation relationship with the spectral components of the first formant and cutoff border areas [27]. However, this method is intended to address band-limited speech and also requires extensive computational expense when applied to full-band speech condition. In this paper, we propose an effective method utilizing the temporal correlation with less computational load for background noise conditions.

A. Missing-Feature Reconstruction Based on Temporal Spectral Feature

The clean speech $X(t)$ at time t in the log-spectral domain (i.e., log-spectral coefficients; logarithm of Mel-filterbank outputs) can be represented by

$$X(t) = [x_1(t), x_2(t), \ldots, x_N(t)]^T \tag{7}$$

where $x_n(t)$ denotes the nth log-spectral component at time frame t and N is the number of log-spectral coefficients which is identical to the number of Mel-filterbanks. Here, we define a *temporal spectral feature* vector of the nth frequency band at time t as

$$X_n^{\{t\}}(t) = [x_n(t-t_d), \ldots, x_n(t), \ldots, x_n(t+t_d)]^T, \quad 1 \le n \le N \tag{8}$$

where t_d denotes a time-lag which determines the analysis range of the temporal spectral feature components. In consequence, the resulting temporal spectral feature vector consists of $2t_d+1$ number of components. The obtained temporal spectral feature vector of clean speech is assumed to be modeled by a Gaussian mixture with $K^{\{t\}}$ components as follows:

$$p\left(X_n^{\{t\}}\right) = \sum_{k=1}^{K^{\{t\}}} \omega_{n,k}^{\{t\}} \mathcal{N}\left(X_n^{\{t\}}; \boldsymbol{\mu}_{n,k}^{\{t\}}, \boldsymbol{\Sigma}_{n,k}^{\{t\}}\right) \tag{9}$$

where $\boldsymbol{\Sigma}_{n,k}^{\{t\}}$ is a full-covariance matrix in a manner similar to (1).

Figs. 1 and 2 illustrate the extraction procedure of the temporal spectral feature vector. The panels of Fig. 1 show clean speech signal in (a) time and (b) frequency (i.e., spectrogram) domains, and (c) the plot of log-spectral components of the 8th band (within 600–700 Hz) of the Mel-filterbanks. The upper plot of Fig. 2 illustrates acquisition processing of the temporal spectral feature vector at each time t_n with a time-lag t_d from the subband log-spectrum signal, which is the beginning part of plot (c) in Fig. 1. The lower figure shows the set of obtained temporal spectral feature vectors for the eighth frequency band from t_1 to t_{11}, resulting in a $(2t_d+1)\times 11$ size matrix. Finally, we obtain N sets of temporal spectral feature vectors over the corresponding N frequency bands for the input speech.

Fig. 3 shows an example of the Gaussian mixture model with eight components obtained by training over the temporal spectral feature of the eighth band, where each plot presents the mean vector of each Gaussian component. In this example, we used six frames for time-lag t_d formulating a 13 dimensional temporal spectral feature vector, which corresponds to 145 ms of time duration.[2] It can be considered that each mean vector represents a pattern of temporal spectral change in log-spectral domain during 145 ms at the eighth frequency band. We found that a relatively small number of Gaussian components $K^{\{t\}}$ are sufficient to represent the statistical patterns of the temporal spectral feature, resulting in less expensive computation compared to an HMM-based method. GMMs with eight components were

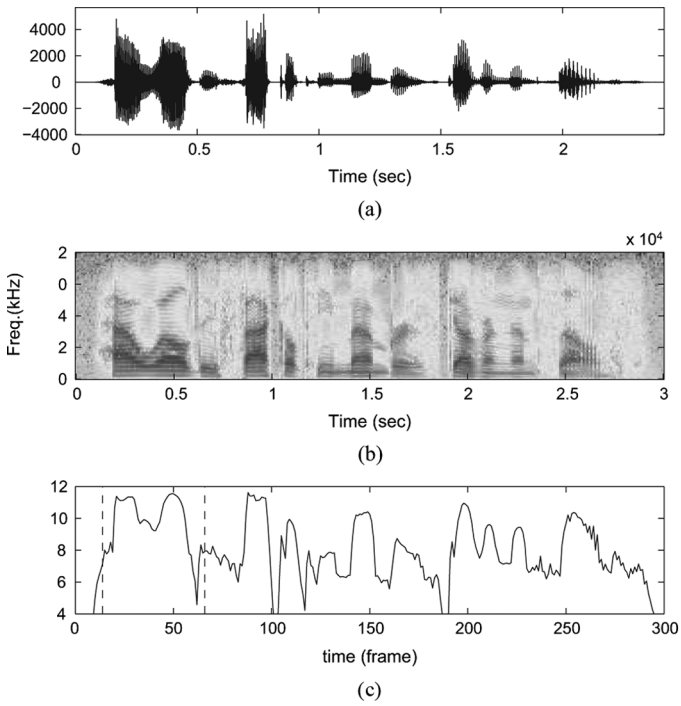


Fig. 1. Clean speech signal in (a) time, (b) frequency (spectrogram), and (c) its log-spectral components of the eighth Mel-frequency band (600-700Hz).

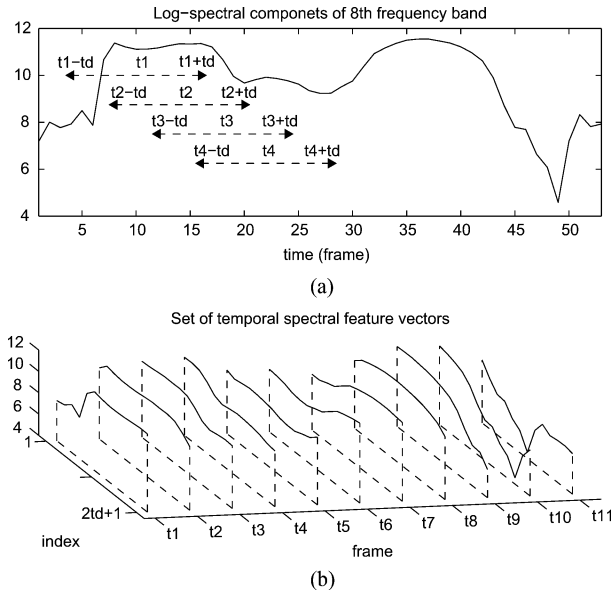


Fig. 2. Temporal spectral feature extraction. (a) Signal acquisition. (b) A set of obtained temporal spectral feature vectors.

[2]Here, a 25 ms analysis window and 10-ms skip rate was used for log-spectral feature extraction in this study (145 ms $= 12 \times 10$ ms $+ 25$ ms).

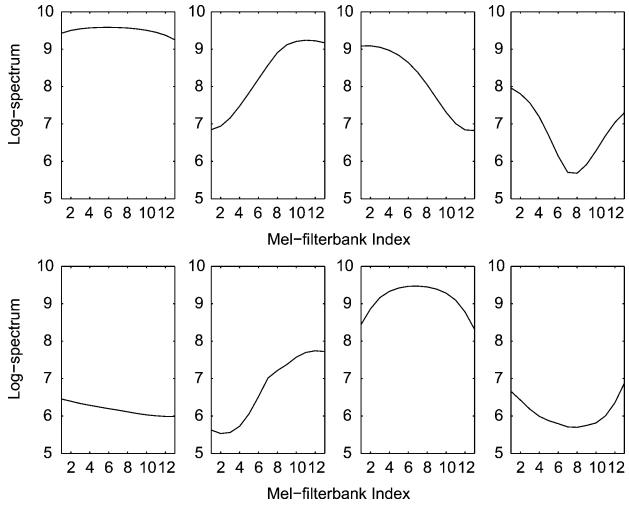


Fig. 3. Mean vectors of a Gaussian mixture model with eight components produced by temporal spectral feature with a time-lag 6.

used for modeling the temporal spectral feature in our experiments.

The feature reconstruction procedure follows the bounded MAP estimation which is employed by the original reconstruction method. The input temporal spectral feature vector $Y_n^{\{t\}}(t)$ of the nth frequency band is considered to consist of reliable $Y_{n,r}^{\{t\}}(t)$ and unreliable components $Y_{n,u}^{\{t\}}(t)$ as follows:

$$\begin{aligned} Y_n^{\{t\}}(t) &= [Y_{n,r}^{\{t\}}(t), Y_{n,u}^{\{t\}}(t)]^T \\ &= [X_{n,r}^{\{t\}}(t), Y_{n,u}^{\{t\}}(t)]^T \end{aligned} \tag{10}$$

where the reliable components $Y_{n,r}^{\{t\}}(t)$ can be replaced with $X_{n,r}^{\{t\}}(t)$ of clean speech. A cluster of the clean speech GMM in (9) is determined by a marginal integration as follows:

$$\hat{k} = \arg\max_k \left\{ P(k) \int_{-\infty}^{Y_{n,u}^{\{t\}}(t)} P(X_n^{\{t\}}(t) \mid k) dX_{n,u}^{\{t\}}(t) \right\}. \tag{11}$$

The missing components $X_{n,u}^{\{t\}}(t)$ of the temporal spectral feature vector is reconstructed by

$$\tilde{X}_{n,u}^{\{t\}}(t) = \boldsymbol{\mu}_{n,\hat{k},u}^{\{t\}} + \boldsymbol{C}_{n,\hat{k},ru}^{\{t\}} \cdot \boldsymbol{C}_{n,\hat{k},rr}^{\{t\}-1} \cdot \left[Y_{n,r}^{\{t\}}(t) - \boldsymbol{\mu}_{n,\hat{k},r}^{\{t\}} \right]. \tag{12}$$

In a manner similar to that for (5) and (6), $\boldsymbol{C}_{n,\hat{k},rr}^{\{t\}}$ and $\boldsymbol{C}_{n,\hat{k},ru}^{\{t\}}$ are defined as follows:

$$\boldsymbol{C}_{n,\hat{k},rr}^{\{t\}} = E\left\{ \left(X_{n,r}^{\{t\}}(t) - \boldsymbol{\mu}_{n,\hat{k},r}^{\{t\}} \right) \left(X_{n,r}^{\{t\}}(t) - \boldsymbol{\mu}_{n,\hat{k},r}^{\{t\}} \right)^T \right\} \tag{13}$$

$$\boldsymbol{C}_{n,\hat{k},ru}^{\{t\}} = E\left\{ \left(X_{n,r}^{\{t\}}(t) - \boldsymbol{\mu}_{n,\hat{k},r}^{\{t\}} \right) \left(X_{n,u}^{\{t\}}(t) - \boldsymbol{\mu}_{n,\hat{k},u}^{\{t\}} \right)^T \right\} \tag{14}$$

where $\boldsymbol{\mu}_{n,\hat{k},r}^{\{t\}}$ and $\boldsymbol{\mu}_{n,\hat{k},u}^{\{t\}}$ are the mean vectors of the $\hat{k}$th cluster of reliable and unreliable components of the clean speech model obtained by (9), respectively.

Different from the conventional reconstruction method presented in Section II, the correlation across neighboring frames is utilized by employing the temporal spectral feature in our proposed method. We name this proposed method *temporal correlation* based missing-feature method in this study. The performance of the proposed method as a change in the time-lag size will be discussed in a later section.

B. Reconstruction by Selective Combination: Temporal-Frequency Correlation Based Method

In this paper, the final estimation for the missing-feature reconstruction is accomplished by a combination of the estimates obtained by the original frequency-based method and the proposed temporal correlation-based method. Here, we denote the estimate obtained by the frequency correlation-based method as $X_u^{\{f\}}(t)$. Through a mask estimation prior to the reconstruction step, we obtain mask information $M(t) = [m_1(t), m_2(t), \ldots, m_N(t)]^T$, which locates the reliable/unreliable components in the log-spectral domain, that is determined as a binary decision (e.g., 1 or 0) in general. In our proposed method, considering the reliability levels of the current frame and the given frequency band within a time-lag, the estimates are selectively decided as shown in (15) at the bottom of the page, where the threshold values for measuring reliability level $\zeta^{\{f\}}$ and $\zeta^{\{t\}}$ are set as follows:

$$\begin{aligned} \zeta^{\{f\}} &= 0.75 \times N \\ \zeta^{\{t\}} &= 0.75 \times (2t_d + 1). \end{aligned} \tag{16}$$

In the proposed method, the original frequency-based and the proposed temporal correlation methods are independently applied to the input speech. In a following step, the reconstructed components of each frame are selectively determined by (15). If the number of reliable components at the current frame is greater than 75% of the log-spectral coefficient dimension N, then the estimated feature component $\tilde{x}_{n,u}^{\{f\}}(t)$ obtained by the frequency correlation-based method is selected. Next, if the number of reliable components within the time-lag $t - t_d$ to $t + t_d$ at the nth frequency band is greater than 75% of total component number of temporal spectral feature $2t_d + 1$, or there is no reliable component at the current frame, then the estimate components reconstructed by the temporal correlation-based method $\tilde{x}_{n,u}^{\{t\}}(t)$ is

$$\tilde{x}_{n,u}(t) = \begin{cases} \tilde{x}_{n,u}^{\{f\}}(t), & \text{if } \sum_n m_n(t) \geq \zeta^{\{f\}} \\ \tilde{x}_{n,u}^{\{t\}}(t), & \text{else if } \sum_{t-t_d}^{t+t_d} m_n(t) \geq \zeta^{\{t\}} \text{ or } \sum_n m_n(t) = 0 \\ \alpha \tilde{x}_{n,u}^{\{f\}}(t) + (1.0 - \alpha) \tilde{x}_{n,u}^{\{t\}}(t), & \text{otherwise} \end{cases} \tag{15}$$

chosen. If these conditions are not satisfied, a weighted summation of both estimates is used for the final estimate. Based on the reconstruction performance of the frequency and temporal correlation-based methods, respectively, the weight α was chosen as 0.7 in this paper.

IV. Experimental Results

A. Experimental Conditions

The TIMIT speech corpus was used for performance evaluation of the proposed method. A total of 4.1 hours of speech (462 speakers, 4620 utterances) were used for training, and 1.5 hours of data (168 speakers, 1680 utterances) were used for test. The training and the test sets do not overlap each other in speakers and uttered sentences. The data was down-sampled to 8 kHz, so that each speech sample contains 4-kHz full-band frequency. In order to evaluate the performance under various types of background noise conditions, noise corrupted test sets were generated by combining clean speech samples with white noise, car noise, speech babble, and background music audio samples. The white noise, car noise, and speech babble samples were obtained from NOISEX92, 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 1680 utterances at three different SNRs: 5, 10, and 15 dB.

We employed SPHINX3 [35] as the HMM-based speech recognizer to obtain recognition accuracy in background noise conditions. Each HMM represents a tri-phone which consists of three states with an eight-component GMM per state, which is tied with 1138 states. The task has 6233 words as the vocabulary, and the trigram language model is adapted on the TIMIT database using a Broadcast News language model as an initial model. A conventional Mel-frequency cepstral coefficient (MFCC) feature front-end is employed in the experiment, which was suggested by the European Telecommunication Standards Institute (ETSI) [36]. An analysis window of 25 ms in 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.

B. Performance of Baseline and Conventional Methods

Performance of the baseline system (no compensation) was examined with comparison to several existing preprocessing algorithms in terms of speech recognition performance. The framework of this study employs a clean condition trained HMM, so we focus only on speech/feature enhancement methods for the performance comparison, and do not consider acoustic model (i.e., HMM) adaptation. Spectral subtraction (SS) [6], [37] combined with cepstral mean normalization (CMN) was selected as one of the conventional algorithms. They represent some 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 noise components are adaptively estimated using the Expectation–Maximization (EM) algorithm over each test utterance [10]. 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 [38]. Table I demonstrates speech recognition performance (i.e., word error rate, WER) of the baseline system and the conventional algorithms on all background noise conditions.

TABLE I
Recognition Performance of Baseline System and Conventional Methods (WER, %)

White Noise	5dB	10dB	15dB	Avg.
Baseline	98.80	92.96	79.41	90.39
SS+CMN	87.35	65.27	43.52	65.38
VTS	90.73	57.47	28.89	59.03
AFE	68.78	42.47	24.90	45.38
Car Noise	5dB	10dB	15dB	Avg.
Baseline	90.72	62.36	32.85	61.98
SS+CMN	66.17	38.27	21.43	41.96
VTS	75.08	39.17	19.98	44.74
AFE	48.20	29.88	20.24	32.77
Speech babble	5dB	10dB	15dB	Avg.
Baseline	81.75	51.34	26.26	53.12
SS+CMN	68.71	37.26	19.87	41.95
VTS	65.15	33.04	17.46	38.55
AFE	50.68	30.72	19.89	33.76
Background Music	5dB	10dB	15dB	Avg.
Baseline	60.14	36.49	20.73	39.12
SS+CMN	46.84	27.99	17.64	30.82
VTS	44.96	26.08	16.15	29.06
AFE	35.77	22.29	16.25	24.77

TABLE II
Recognition Performance of Conventional Missing-Feature Reconstruction With Oracle Mask (WER, %)

F-MFR with Oracle	5dB	10dB	15dB	Avg.
White Noise	62.11	50.29	40.95	51.12
Car Noise	51.67	34.13	23.67	36.49
Speech Babble	43.66	29.20	19.98	30.95
Background Music	25.51	19.01	14.57	19.70
F-MFR+SS with Oracle	5dB	10dB	15dB	Avg.
White Noise	58.26	46.07	35.31	46.55
Car Noise	40.06	26.58	19.55	28.73
Speech Babble	33.66	22.26	16.24	24.05
Background Music	22.12	17.31	13.29	17.57

Here, we obtained 61.15%, 45.03%, 42.85%, and 34.17% for baseline (no processing), SS + CMN, VTS, and AFE as average WERs over 5, 10, and 15 dB SNRs of all four noise conditions (See Table VI). For comparison of the baseline performance, 37.70%, 18.14%, 16.70%, and 10.48% WERs were obtained by the same processing methods respectively as average performance over 5-, 10-, and 15-dB SNR conditions for the Aurora 2.0 task [14], which is widely used in the research community targeting a small task of connected digits recognition [39]. Therefore, the TIMIT corpus task employed in our experiment has a more challenging configuration in lexical and grammar for speech recognition.

Table II shows recognition performance obtained using the original cluster-based missing-feature reconstruction method with the "Oracle" mask. The oracle mask was generated by comparing the noise-corrupted speech signal to the original clean speech data at the log-spectrum level. Here, we denote the original reconstruction method as F-MFR where "frequency correlation" is employed. In the missing-feature reconstruction of all our experiments, 23rd-order of log-spectral coefficients

(i.e., log of Mel-filterbank output) were used as the feature vector and a 128-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. We found that the missing-feature reconstruction method produces a significant improvement in WER when combined with conventional spectral subtraction (SS). Therefore, in this paper, the performance of the missing-feature methods are evaluated also for the case of a combination with spectral subtraction. It can be seen that the performance of F-MFR with the oracle mask outperforms most other conventional speech/feature enhance methods shown in Table I, in particular, when combined with spectral subtraction (F-MFR+SS). It confirms that the missing-feature reconstruction method shows significant effective performance for speech recognition in adverse background conditions, if the knowledge on the reliable spectral components (i.e., mask information) is properly provided as the oracle mask information is available here. In our paper, the performance of F-MFR is compared, as a baseline performance, to the proposed temporal spectral feature based reconstruction method, which will be evaluated in the next section.

C. Missing-Feature Reconstruction Employing Temporal Spectral Feature Analysis

In this section, the proposed reconstruction method employing the temporal spectral feature is evaluated. First, we observe the impact of the time-lag for the temporal spectral feature analysis on the performance of missing-feature reconstruction. Fig. 4 shows speech recognition performance of the reconstructed feature using the temporal correlation-based method (T-MFR) as a function of time-lag t_d from 2 to 10 frames. Here, temporal correlation is solely used for missing-feature reconstruction with oracle mask along with spectral subtraction, and the WER is an average value over all three SNR conditions. Using the $2t_d + 1$-dimensional feature vector, the time-lags 2 to 10 correspond to an 65 to 225-ms time range for the temporal spectral change. GMMs with eight Gaussian components [i.e., $K^{\{t\}} = 8$ in (9)] were used for acoustic models for temporal spectral feature at each frequency band. It is seen that performance increases (i.e., WER decreases) as we increase the time-lag size and performance levels off after around time-lag 6. The results also demonstrate that the performance of the temporal correlation-based method (solid line with black faced circles) is inferior to the original frequency correlation based method (dashed line) in average WER.

Fig. 5 presents performance plots of the combination method of frequency (F-MFR) and temporal (T-MFR) correlation methods, which is denoted as TF-MFR in this paper. In this evaluation, we also see similar performance trends compared to plots in Fig. 4, with considerable improvement in the average performance of the combination method (TF-MFR) compared to the original reconstruction method (F-MFR). From our series of experiments, it was found that the performance of TF-MFR consistently drops for all noise conditions as the time-lag increases to 8 and 10, in the case of combination with CMN. We conclude that a suitable length of time-lag is required to be

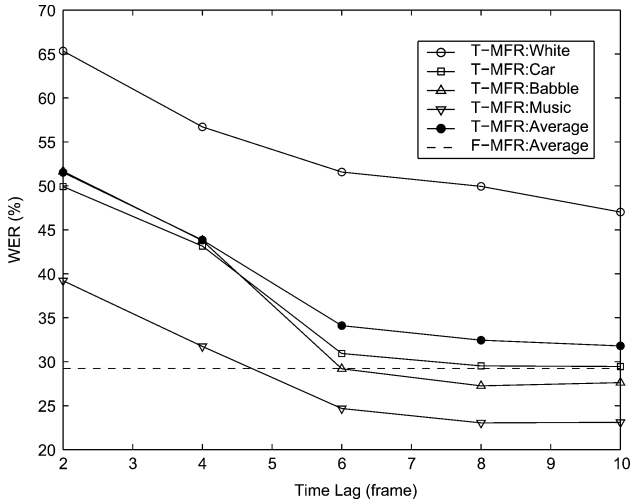


Fig. 4. Recognition performance of the temporal correlation based method (T-MFR) as change of time-lag.

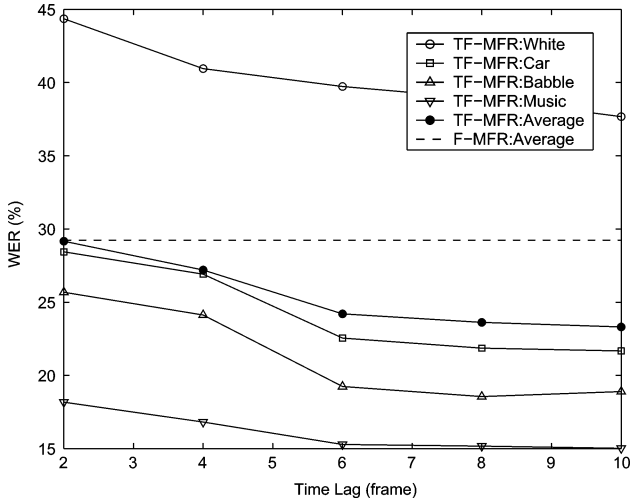


Fig. 5. Recognition performance of the combination of temporal and frequency correlation-based methods (TF-MFR) as change of time-lag.

selected for performance improvement when utilizing temporal correlation. Fig. 6 compares examples of the mean vectors of GMMs generated by different length time-lags (e.g., 2, 6, and 10) for the temporal spectral feature extraction. It can be considered that when the time-lag is too short, it is not effective to reconstruct the feature, since sufficient amount of knowledge on reliable components cannot be provided. An excessively large size time-lag would also produce incorrect estimates, if the number of reliable components is relatively small compared to the size of the temporal spectral feature vector. We used six frames of the time-lag for the temporal spectral feature analysis in the following experiments for the best performance.

Tables III and IV show performance of the proposed temporal correlation-based (T-MFR{+SS}) and combination methods (TF-MFR{+SS}) for all background noise conditions, when solely used (Table III) and combined with spectral subtraction (Table IV), respectively. The relative improvement is calculated compared to the original frequency correlation-based method (F-MFR{+SS}). These results confirm that the leveraged temporal correlation is considerably effective at increasing recognition performance of the reconstructed feature which is obtained

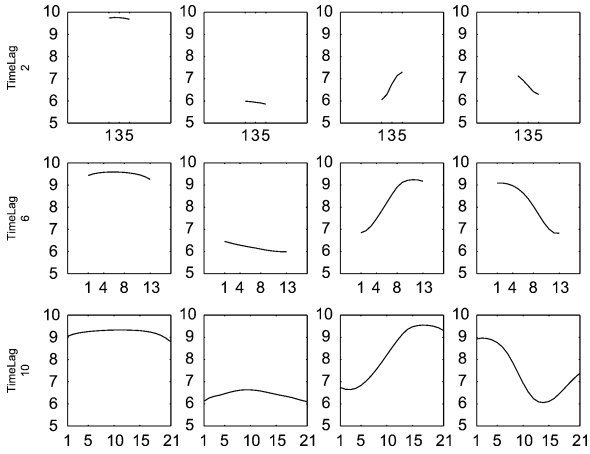


Fig. 6. Examples of mean vectors obtained by different time-lags (a) 2, (b) 6, and (c) 10 frames for temporal spectral feature analysis.

TABLE III
RECOGNITION PERFORMANCE IN WER (%) OF PROPOSED MISSING-FEATURE RECONSTRUCTION EMPLOYING TEMPORAL CORRELATION WITH ORACLE MASK: RELATIVE IMPROVEMENT OF TF-MFR COMPARED TO F-MFR IS SHOWN IN A PARENTHESIS $(+\mathrm{X.XX}\%)$

White Noise	5dB	10dB	15dB	Avg.
F-MFR	62.11	50.29	40.95	51.12
T-MFR	86.89	66.58	45.72	66.40
TF-MFR	**62.17 (-0.10)**	**44.84 (+10.84)**	**33.88 (+17.26)**	**46.96 (+9.34)**
Car Noise	5dB	10dB	15dB	Avg.
F-MFR	51.67	34.13	23.67	36.49
T-MFR	80.11	49.03	25.28	51.47
TF-MFR	**49.80 (+3.62)**	**29.17 (+14.53)**	**17.33 (+26.78)**	**32.10 (+14.98)**
Speech Babble	5dB	10dB	15dB	Avg.
F-MFR	43.66	29.20	19.98	30.95
T-MFR	70.92	41.52	21.10	44.51
TF-MFR	**40.64 (+6.92)**	**23.93 (+18.05)**	**14.83 (+25.78)**	**26.47 (+16.91)**
Background Music	5dB	10dB	15dB	Avg.
F-MFR	25.51	19.01	14.57	19.70
T-MFR	45.45	27.40	17.03	29.96
TF-MFR	**23.27 (+8.78)**	**16.37 (+13.89)**	**11.84 (+18.74)**	**17.16 (+13.80)**

only by the frequency correlation-based method, across all noise conditions in types and SNRs. The performance comparison of the proposed temporal-frequency correlation-based missing-feature reconstruction method $(\text{TF-MFR}\{\text{+SS}\})$ to the original frequency correlation method $(\text{F-MFR}\{\text{+SS}\})$ is summarized in Table V as average WERs over the four types of background noise conditions. By leveraging the temporal correlation, we obtained $+13.76\%$ and $+17.71\%$ average relative improvements[3] in WER for all noise conditions when solely used and combined with spectral subtraction, respectively, compared to the original frequency correlation-based method.

D. Performance Evaluation Employing Mask Estimation Method

Although mask estimation is not in the scope of our study for this paper, the evaluation employing a mask estimation method

[3]The average relative improvement is computed by taking the average of the obtained relative improvements (i.e., $13.76 = (4.81 + 14.33 + 22.14)/3$).

TABLE IV
RECOGNITION PERFORMANCE IN WER (%) OF PROPOSED MISSING-FEATURE RECONSTRUCTION EMPLOYING TEMPORAL CORRELATION WITH ORACLE MASK AND SPECTRAL SUBTRACTION: RELATIVE IMPROVEMENT OF TF-MFR COMPARED TO F-MFR IS SHOWN IN A PARENTHESIS $(+\mathrm{X.XX}\%)$

White Noise	5dB	10dB	15dB	Avg.
F-MFR+SS	58.26	46.07	35.31	46.55
T-MFR+SS	70.01	48.90	35.82	51.58
TF-MFR+SS	**52.00 (+10.74)**	**38.46 (+16.52)**	**28.72 (+18.66)**	**39.73 (+15.31)**
Car Noise	5dB	10dB	15dB	Avg.
F-MFR+SS	40.06	26.58	19.55	28.73
T-MFR+SS	48.67	26.80	17.36	30.94
TF-MFR+SS	**33.51 (+16.35)**	**19.91 (+25.09)**	**14.24 (+27.16)**	**22.55 (+22.87)**
Speech Babble	5dB	10dB	15dB	Avg.
F-MFR+SS	33.66	22.26	16.24	24.05
T-MFR+SS	43.93	25.92	17.76	29.20
TF-MFR+SS	**27.11 (+19.46)**	**17.07 (+23.32)**	**13.55 (+16.56)**	**19.24 (+19.78)**
Background Music	5dB	10dB	15dB	Avg.
F-MFR+SS	22.12	17.31	13.29	17.57
T-MFR+SS	34.30	22.65	17.10	24.68
TF-MFR+SS	**19.25 (+12.97)**	**14.74 (+14.85)**	**11.85 (+10.84)**	**15.28 (+12.89)**

TABLE V
PERFORMANCE COMPARISON IN WER (%) IN ALL SNR CONDITIONS AS AVERAGE OVER FOUR TYPES OF BACKGROUND NOISE CONDITIONS, WHERE F-MFR AND TF-MFR ARE WITH ORACLE MASK: RELATIVE IMPROVEMENT OF TF-MFR COMPARED TO F-MFR IS SHOWN IN A PARENTHESIS $(+\mathrm{X.XX}\%)$

	5dB	10dB	15dB	Avg.
F-MFR	45.74	33.16	24.79	34.56
TF-MFR	**43.97 (+4.81)**	**28.58 (+14.33)**	**19.47 (+22.14)**	**30.67 (+13.76)**
F-MFR+SS	38.53	28.06	21.10	29.23
TF-MFR+SS	**32.97 (+14.88)**	**22.55 (+19.94)**	**17.09 (+18.31)**	**24.20 (+17.71)**

(without Oracle knowledge) would suggest an available performance of the proposed method when applied with an actual mask estimation technique in real-life scenarios. Here, we employed a mask estimation method which utilizes a Posterior-based *Representative* Mean (PRM) estimate for determining the reliability of the input speech spectrum, that has been proposed in our recent study [40]. In this method, the PRM estimate 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 was employed, which was previously proposed for feature compensation [14]. The mask of the mth frequency band at time t is determined by comparing the ratio of the mth PRM estimates of noise-corrupted and clean speech in log-spectral domain as follows:

$$\frac{\tilde{\mu}_Y(t,m)}{\tilde{\mu}_X(t,m)} \underset{\text{reliable}}{\overset{\text{unreliable}}{\gtrless}} \zeta_{\text{PRM}} \tag{17}$$

where the threshold ζ_{PRM} was set to 1.15 for all noise conditions in our experiment.

Employing the PRM-based mask estimation, the proposed temporal-frequency correlation based method was compared to the original F-MFR and other conventional preprocessing

TABLE VI
PERFORMANCE COMPARISON IN WER (%) IN FOUR TYPES OF BACKGROUND NOISE CONDITIONS AS AVERAGE OVER ALL SNRS; 5, 10, 15 dB, WHERE F-MFR AND TF-MFR ARE WITH PRM-BASED MASK ESTIMATOR: RELATIVE IMPROVEMENT OF TF-MFR COMPARED TO F-MFR IS SHOWN IN A PARENTHESIS ($+$X.XX%)

	White	Car	Babble	Music	Avg.
Baseline	90.39	61.98	53.12	39.12	61.15
SS+CMN	65.38	41.96	41.95	30.82	45.03
VTS	59.03	44.74	38.55	29.06	42.85
VTS+SS+CMN	50.84	37.07	38.57	29.47	38.99
AFE	45.38	32.77	33.76	24.77	34.17
F-MFR+SS	61.44	41.65	39.31	32.99	43.85
TF-MFR+SS	**61.93 (-0.82)**	**39.16 (+8.33)**	**37.78 (+5.83)**	**30.97 (+7.61)**	**42.46 (+5.24)**
F-MFR+SS+CMN	55.31	41.19	38.98	30.74	41.55
TF-MFR+SS+CMN	**53.62 (+3.67)**	**37.88 (+10.27)**	**37.04 (+6.95)**	**28.49 (+9.09)**	**39.26 (+7.50)**

TABLE VII
PERFORMANCE COMPARISON IN WER (%) IN ALL SNRS CONDITIONS AS AVERAGE OVER FOUR TYPES OF BACKGROUND NOISE CONDITIONS, WHERE F-MFR AND TF-MFR ARE WITH PRM-BASED MASK ESTIMATOR: RELATIVE IMPROVEMENT OF TF-MFR COMPARED TO F-MFR IS SHOWN IN A PARENTHESIS ($+$X.XX%)

	5dB	10dB	15dB	Avg.
Baseline	82.85	60.79	39.81	61.15
SS+CMN	67.27	42.20	25.62	45.03
VTS	68.98	38.94	20.62	42.85
VTS+SS+CMN	59.19	35.60	22.17	38.99
AFE	50.86	31.34	20.32	34.17
F-MFR+SS	62.78	41.25	27.51	43.85
TF-MFR+SS	**62.70 (+0.33)**	**39.20 (+5.80)**	**25.48 (+9.59)**	**42.46 (+5.24)**
F-MFR+SS+CMN	62.27	38.35	24.04	41.55
TF-MFR+SS+CMN	**60.94 (+2.24)**	**35.63 (+7.63)**	**21.20 (+12.62)**	**39.26 (+7.50)**

methods in Tables VI and VII. From the results, the proposed TF-MFR method showed a $+5.24\%$ average relative improvement in WER for all four background noise conditions, compared to the F-MFR, when combined with spectral subtraction. It was found that combination with CMN more increases the performance of the proposed TF-MFR. In case of combination with CMN, the TF-MFR showed 39.26% in average WER which is comparable to performance of VTS+SS+CMN (38.99%), presenting $+7.50\%$ relative improvement compared to the F-MFR. These results confirm that the proposed method could be effective at more increasing speech recognition performance in various adverse background noise conditions where mask information is unknown, by employing an effective mask estimator. It can be seen that WER of TF-MFR with the PRM-based mask estimator is still higher compared to the Oracle case (Tables III–V) and AFE[4] (34.17%); however, this can be addressed to some extent in the future by employing a more effective mask estimation technique. In our study [40], the missing-feature method with the PRM-based mask estimator outperformed both VTS and AFE for the speech babble and music noise conditions with the Aurora 2.0 task. It is considered that more elaborated mask estimation method is required to more improve performance of the proposed TF-MFR method on the TIMIT corpus, by addressing the larger complexity of speech structure included in the TIMIT database, compared to the Aurora 2.0 which is a connected-digit task. Next, the evaluation moves to actual in-vehicle speech data.

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

The proposed temporal-frequency correlation-based method was also evaluated on a real-life in-vehicle condition obtained from the CU-Move corpus [2]. 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, produced over 600 GB of data during a six-month collection effort across the U.S. The database and noise conditions are discussed in detail in [2]. 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 [41].

Table VIII shows the performance evaluation of the proposed TF-MFR on the CU-Move corpus. Here, we employed the identical PRM-based mask estimation as presented by (17), where the posterior-based representative mean estimates are compared to a threshold for binary decision. These results demonstrate that TF-MFR+SS brings consistent improvement compared to the original F-MFR+SS on the real-life in-vehicle condition as well, resulting in $+4.56\%$ and $+16.72\%$ relative improvements solely used and combined with CMN, respectively. It is noted that the proposed TF-MFR+SS combined with CMN significantly outperforms the sole TF-MFR+SS ($32.62\% \rightarrow 28.30\%$), while the original F-MFR+SS+CMN is slightly better than the F-MFR+SS ($34.18\% \rightarrow 33.98\%$). We believe that the TF-MFR generates a more accurate spectral contour in the time domain by utilizing temporal correlation, having CMN more effectively estimate convolutional noise components which would be found in actual in-vehicle environments. The results also show that the performance of the proposed TF-MFR+SS+CMN is considerably more effective on the CU-Move corpus, compared to SS+CMN, VTS, and AFE. The proposed TF-MFR+SS+CMN is still worse than VTS+SS+CMN in this experiment, however, the performance difference could be compensated with a more effective mask estimator. The results here also prove that the proposed TF-MFR method could be applicable to real-life in-vehicle conditions to improve performance of speech recognition.

V. CONCLUSION

In this paper, a missing-feature reconstruction method was proposed to improve speech recognition performance in various types of background noise environments. The conventional cluster-based missing-feature reconstruction method utilizes

[4]AFE showed the best performance without SS or CMN.

[5]0-dB and 5-dB SNR test samples of the car noise condition of Aurora2.0 show 7.15-dB and 11.66-dB average SNRs, respectively, using the NIST tool.

TABLE VIII
RECOGNITION PERFORMANCE IN WER (%) COMPARISON FOR THE CU-MOVE CORPUS WITH PRM-BASED MASK ESTIMATION: RELATIVE IMPROVEMENT COMPARED TO F-MFR IS SHOWN IN A PARENTHESIS ($+$X.XX%)

Baseline	70.02	
SS+CMN	39.90	
VTS	48.31	
VTS+SS+CMN	23.29	
AFE	31.45	
F-MFR+SS	34.18	
TF-MFR+SS	**32.62**	**(+4.56)**
F-MFR+SS+CMN	33.98	
TF-MFR+SS+CMN	**28.30**	**(+16.72)**

only log-spectral correlation across frequency bands. To increase performance of the missing-feature reconstruction by leveraging temporal correlation across neighboring frames, temporal spectral feature analysis was developed which uses the log-spectral coefficients with a time-lag. In a manner similar with the original reconstruction method, a Gaussian mixture model was obtained by training on the extracted temporal spectral feature set and a bounded MAP estimation was used for feature reconstruction. The missing-feature was finally reconstructed by a selective combination of the original frequency correlation-based method and the proposed temporal correlation-based method.

The performance of the proposed method was evaluated on the TIMIT speech corpus using various types of additive background noise conditions and the CU-Move actual in-vehicle corpus. Experimental results demonstrated that the proposed method is more effective at increasing speech recognition performance in adverse conditions. A suitable size time-lag for the temporal spectral feature extraction was needed for improved reconstruction performance. By employing the proposed temporal-frequency correlation based reconstruction method with oracle mask information, we obtained a $+17.71\%$ average relative improvement in WER for white, car, speech babble, and background music conditions over 5-, 10-, and 15-dB SNR, compared to the original frequency correlation based method. The proposed method also obtained a $+16.72\%$ relative improvement employing an actual mask estimation in real-life in-vehicle conditions using CU-Move data.

REFERENCES

[1] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. H. L. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *IEEE Intell. Vehicle Conf.*, 2007.

[2] 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 *Chap. 2 in DSP for In-Vehicle and Mobile Systems*, Abut, Hansen, and Takeda, Eds. New York: Springer, 2004.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Commun.*, vol. 37, pp. 231–248, 2002.

[12] V. Stouten, H. Van hamme, and P. Wambacq, "Joint removal of additive and convolutional noise with model-based feature enhancement," in *Proc. ICASSP2004*, 2004, pp. 949–952.

[13] A. Sasou, T. Tanaka, S. Nakamura, and F. Asano, "HMM-based feature compensation methods: an evaluation using the Aurora2," in *Proc. ICSLP2004*, 2004, pp. 121–124.

[14] 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.

[15] 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, 1994.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] H. Van Hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP'04*, May 2004, pp. 213–216.

[23] 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.

[24] 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.

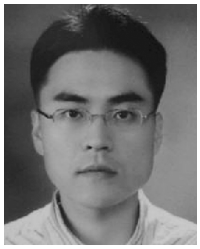
[25] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. ICASSP'06*, May 2006, pp. 305–308.

[26] 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.

[27] 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.

[28] 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, Sep. 2004.

[29] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 2000.
[30] B. J. Borgstrom and A. Alwan, "HMM-based estimation of unreliable spectral components for noise robust speech recognition," in *Proc. Interspeech'08*, 2008, pp. 1769–1772.
[31] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[32] H. Hermansky and S. Sharma, "Temporal patters (TRAPS) in ASR of noisy speech," in *Proc. ICASSP'99*, 1999, pp. 289–292.
[33] C.-P. Chen, J. Bilmes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on Aurora 2.0," in *Proc. ICSLP'02*, 2002, pp. 2445–2448.
[34] X. Xiao, E.-S. Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1662–1674, Nov. 2008.
[35] [Online]. Available: http://cmusphinx.sourceforge.net
[36] *ETSI Standard Document*, ETSI ES 201 108 v1.1.2 (2000-04), 2000.
[37] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO-94*, 1994, pp. 1182–1185.
[38] *Etsi Standard Document*, ETSI ES 202 050 v1.1.1 (2002-10), 2002.
[39] 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.
[40] 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*, Dec. 2009, pp. 194–198.
[41] "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, Korea, 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 (2005–2007) and a Postdoctoral Researcher in the electrical and computer engineering, Carnegie Mellon University, Pittsburgh, PA (2004–2005), and Korea University (2003–2004), respectively. 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 Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, NJ, in 1982.

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, Boulder, (1998–2005), where he cofounded 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 50 (22 Ph.D., 28 M.S./M.A.) thesis candidates, was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body, author/coauthor of 359 journal and conference papers and eight textbooks in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), 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 for contributions in "Robust Speech Recognition in Stress and Noise," in 2007 and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2009–2011; 2006–2008) and Educational Technical Committee (2006–2008; 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 guest editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS 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 International Speech Communications Association (ISCA) Advisory Council. He also organized and served as General Chair for Interspeech-2002/ICSLP-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and served as Technical Program Chair and Co-Organizer for the IEEE ICASSP-2010, Dallas, TX.