Feature Compensation Employing Multiple Environmental Models for Robust In-Vehicle Speech Recognition

Wooil KIM†, Nonmember and John H.L. HANSEN†(a), Member

SUMMARY An effective feature compensation method is developed for reliable speech recognition in real-life in-vehicle environments. The CU-Move corpus, used for evaluation, contains a range of speech and noise signals collected for a number of speakers under actual driving conditions. The authors employ the Environment Transition Model for mixture sharing. The proposed scheme eliminates the prior training which requires a noise-corrupted speech database, which is an absolute requirement in conventional data-driven methods such as RATZ [14] and SPLICE [15]. The corpus has been used for research in multi-sensor array processing for noise suppression and speech recognition in cars [12]. Therefore, performance evaluation on the CU-Move corpus can indicate the reliability and effectiveness of the targeted algorithm in actual in-vehicle conditions.

In this study, our previously proposed PCGMM (Parallel Combined Gaussian Mixture Model) based feature compensation method [13] is considered as a solution to address the background noise of in-vehicle conditions. PCGMM-based method employs model combination for noise-corrupted speech model and operates in the cepstral domain. By using model combination, the PCGMM scheme eliminates the prior training which requires a noise-corrupted speech database, which is an absolute requirement in conventional data-driven methods such as RATZ [14] and SPLICE [15].

In this paper, we especially focus on the PCGMM method employing multiple environmental models [16]. Multiple model approach no longer requires a separate algorithm for silence detection which is indispensable in the single model approach for PCGMM. In order to reduce the computational expense due to the use of multiple models, we employ an Environment Transition Model for the multi-model approach, which is motivated from the Noise Language Model in our previous work [17], [18]. To take test data to be closer to the original training condition by compensating the speech signal or extracted features. Alternatively, the second category would concentrate on transforming the prior trained acoustic model to be closer to the test speech acoustics. Speech enhancement, feature processing such as Cepstral Mean Normalization (CMN), and many types of feature compensation methods are examples of the first category [1]–[6]. Methods belonging to the second category are not directed at removing noise components, but generating a speech model which matches better the noisy environment during the training or decoding steps. The Maximum A Posteriori (MAP) [7], [8] and Maximum Likelihood Linear Regression (MLLR) [9], as well as EigMap [10] adaptation techniques are included in this category.

This paper investigates the performance of our feature compensation scheme in a real-life in-vehicle environment, with the goal of achieving robust performance with lower computational expenses. The CU-Move corpus has been built to develop reliable speech systems for in-vehicle and it contains a range of acoustic signals expected to be observed during real-life car-driving [11]. The corpus has been used for research in multi-sensor array processing for noise suppression and speech recognition in cars [12]. Therefore, performance evaluation on the CU-Move corpus can indicate the reliability and effectiveness of the targeted algorithm in actual in-vehicle conditions.

In this study, our previously proposed PCGMM (Parallel Combined Gaussian Mixture Model) based feature compensation method [13] is considered as a solution to address the background noise of in-vehicle conditions. PCGMM-based method employs model combination for noise-corrupted speech model and operates in the cepstral domain. By using model combination, the PCGMM scheme eliminates the prior training which requires a noise-corrupted speech database, which is an absolute requirement in conventional data-driven methods such as RATZ [14] and SPLICE [15].

In this paper, we especially focus on the PCGMM method employing multiple environmental models [16]. Multiple model approach no longer requires a separate algorithm for silence detection which is indispensable in the single model approach for PCGMM. In order to reduce the computational expense due to the use of multiple models, we employ an Environment Transition Model for the multi-model approach, which is motivated from the Noise Language Model in our previous work [17], [18]. To take
advantage of the environment transition model, an environment dependent scheme for mixture sharing is proposed in this paper.

This paper is organized as follows. We first review the CU-Move corpus used for this study in Sect. 2. In Sect. 3, PCGMM-based feature compensation employed in our work will be discussed, followed by the multi-model approach for PCGMM in Sect. 4. We also discuss the environment transition model and the environment dependent implementation of the mixture sharing technique in Sect. 5 and Sect. 6 respectively. Representative experimental procedures and their results are presented and discussed in Sect. 7. Finally, in Sect. 8, we conclude our work.

2. CU-Move Corpus

The CU-Move project[11] 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; (i) command and control words, (ii) digit strings of telephone and credit numbers, (iii) street names and addresses, (iv) phonetically-balanced sentences, and (v) Wizard of Oz interactive navigation conversations. A total of 500 speakers, balanced across gender and age, produced over 600GB of data during a six-month collection effort across the United States. Figure 1 shows the vehicle setup for data collection of CU-Move corpus. The database and noise conditions are discussed in detail in [11]. We point out that the noise conditions are changing with time and are quite different in terms of SNR, stationarity and spectral structure. The challenge in addressing these noise conditions is that they might be changing depending on the car being used and the road. In this study, we selected 20 speakers from approximately 100 speakers in Minn., MN (i.e., Release 1.1A) and employ the connected single digits portion that contains speech under a range of varying complex in-vehicle noise events/conditions.

3. PCGMM-Based Feature Compensation

In the PCGMM-based method, the parameters of the noise-corrupted speech model are obtained through a model combination procedure using clean speech and noise models independently[13]. A constant bias transformation of the mean parameters of the clean speech model is assumed in the cepstral domain under the additive noisy environment, which is the assumption generally taken by other data-driven methods[14],[19] as follows,

$$\hat{\mu}_{y,k} = \mu_{x,k} + r_k.$$  (1)

where $\mu_{y,k}$ and $\mu_{x,k}$ denote mean vectors of $k$th component of GMMs for noise corrupted speech $y$ and clean speech $x$ respectively. The bias term $r_k$ is used for reconstruction of the input speech. The bias term is estimated with Eq. (1), once the mean parameters of the clean speech model and corresponding noise-corrupted speech model are obtained.

The MMSE equation for reconstruction of the clean speech is approximated using Eq. (2) as follows[19],

$$\hat{s}_{MMSE} = \int_{X} x p(x|y) dx = y - \sum_{k=1}^{K} r_k p(k|y).$$  (2)

The posterior probability $p(k|y)$ can be calculated based on the parameters of the noisy speech GMM $\{\omega_k, \mu_{y,k}, \sigma_{y,k}\}$ which consists of $K$ number of Gaussian components.

4. PCGMM-Based Method Employing Multiple Environmental Models

In the PCGMM-based method presented in Sect. 3 where a single model is employed, model adaptation needs to be applied in order to address the time-varying background noise. In the current framework for PCGMM-based method, the noise model is updated during silence periods via adaptation followed by combination of models, which again reflects the true noise for the GMM of the noisy speech. Such a framework, however, requires an accurate algorithm for silence detection and also needs considerable computational resources due to the conversion between the linear spectrum, log spectrum and cepstral domains. Therefore, employing the PCGMM-based method with a single model may not be appropriate for some particular applications requiring fast (or real-time) processing. In this section, we consider the PCGMM-based method that employs a combination of multiple environmental models for real-time processing.

Utilizing a multiple number of environmental models obtained off-line can be considered to be effective for compensating input features adaptively under time-varying noisy conditions and eliminating the need for additional silence detection and online model combination. In a multiple model method, the posterior probability of each possible environment is estimated over the incoming noisy speech. In our work, the feature reconstruction procedure is modified using a frame-by-frame formulation for real-time processing by defining the sequential posterior probability of the environment[16]. Given the incoming noisy speech feature
vectors $Y_t = \begin{bmatrix} y_1, y_2, \ldots, y_t \end{bmatrix}^T$, the sequential posterior probability of a specific environment GMM $G_i$ among models over the input speech feature $Y_t$ can be written as,

$$p(G_i | Y_t) = \frac{p(G_i) p(Y_{t-1} | G_i) p(y_t | G_i)}{\sum_{i=1}^{E} p(G_e) p(Y_{t-1} | G_e) p(y_t | G_e)}, \quad (3)$$

where $p(Y_{t-1} | G_e) = \prod_{t=1}^{t-1} p(y_t | G_e)$ and $p(G_i)$ is a prior probability of each environment represented as a GMM. Based on Eq. (3), the clean feature at frame $t$ is reconstructed by the weighted combination of the compensation terms obtained from a set of $E$ multiple environments as follows,

$$\hat{x}_t, \text{MMSE} \equiv y_t - \sum_{e=1}^{E} p(G_e | Y_t) \sum_{k=1}^{K} r_{e,k} p(k | G_e, y_t), \quad (4)$$

where $r_{e,k}$ is a constant bias term from the $k$th Gaussian component of the $e$th environment model and $p(k | G_e, y_t)$ is the posterior probability for environment $G_e$. Figure 2 demonstrates the multiple model approach of PCGMM-based feature compensation method. Each environmental model $G_e$ is estimated through the model combination procedure using the clean speech GMM and each different noise model which is obtained from off-line training. When the background noise comes from an environment where the number of unique types is finite, such as for in-vehicle conditions (e.g., engine noise, wind noise, turn signal noise, wiper blade noise, etc. [18]) which are considered in this paper, the multiple-model method would be more effective than adaptation or online estimation of noise components in terms of computational complexity.

5. Environment Transition Model

The computational expense for model-based feature compensation depends primarily on the number of Gaussian components to be considered for calculating the probability values (scores). It is true that the computational complexity increases in proportion to the number of environmental models employed for the multi-model approach described in Sect. 4. However, more accurate modeling for noisy conditions requires a larger number of GMMs with sufficient sized pdfs. Now, we introduce an Environment Transition Model in an effort to reduce the computational complexity in the multi-model approach. The motivation is that there might be a smaller sized set of noise types among all types of noise which we need to consider at a certain time frame or session when employing multiple environmental models for PCGMM-based feature compensation.

The environment transition model is motivated from the Noise Language Model which was employed in the Environmental Sniffing scheme developed to decode the most likely sequence of noise types [18]. In order to build the noise language model, in-vehicle acoustic data (i.e., a Blazer SUV) was collected during a 17-mile driving route which contains samples of all driving conditions expected in city and rural areas, and then the primary noise conditions were identified as follows:

- **N1**: idle noise, no movement, windows closed
- **N2**: city driving without traffic, windows closed
- **N3**: city driving with traffic, windows closed
- **N4**: highway driving, windows closed
- **N5**: highway driving, windows 2 inches open
- **N6**: highway driving, windows half-way down
- **N7**: windows 2 inches open in city traffic
- **NX**: others

A total of 3 hours of noise samples were used to build the Noise Language Model, with each segmented noise of length varying from a few seconds to 30 seconds. The analysis of the classified noise samples shows that they have different characteristics in energy levels and statistical based spectral structure [20], [21]. A bigram type of noise language model was constructed using the CMU-Cambridge Statistical Language Modeling (SLM) Toolkit. From the Noise Language Model, only the absolute connectivity between the noise conditions is employed for the environment transition model, since we do not encode specific transition probabilities [17]. Figure 3 shows the environment transition model employed in this paper. The connectivity of the environment transition model in Fig. 3 can be written in matrix notation as follows,

$$C_{ETM} = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix} \quad (5)$$
where each index corresponds to each environment (i.e., noise condition, N1, ..., N7) and NX is considered as the 8th element. Each element in \( C_{ETM}(i, j) \) can be written as \( c_{ij} \), and considered as prior knowledge on the \( j \)th environment for multiple environmental models in the current speech (or frame, session) when the \( i \)th environment is determined as the most likely observed condition in the previous speech.

The environment transition model is applied to the current input speech based on the type of environment most likely observed at the previous utterance, assuming that the current speech was produced in a continuing driving condition from the previous utterance. The most likely observed environment is determined by the maximum of the posterior probability of an environment \( p(G_i|Y) \) over the input speech \( Y \). Suppose that the N4 condition (highway driving, windows closed) was determined at the previous utterance. Only four noise conditions (N3, N4, N5, and NX) are considered for multiple environmental models by replacing the prior probabilities \( P(G_i) \) in Eq. (3) with the normalized connectivity \( c_{ij} / \sum_j c_{ij} \) of the environment transition model in Eq. (5). This case is expected to result in a reduced computational expense (= 4/9 amount) compared to the case of not employing the transition model (i.e., fully connected environmental models).

### 6. Environment Dependent Mixture Sharing Scheme

In this section, we describe another scheme to more effectively reduce the computational expense in the multiple model approach. In our scheme, statistically similar Gaussian components across the different environmental models are selected and the common components for sharing are generated by merging the similar components [16]. The Kullback-Leibler distance is used to represent the statistical similarity between Gaussian components of the GMM models. In our previous work, the mixture sharing technique was applied across all environmental models [16], which is named the environment independent style here. Increasing the number of shared Gaussian components for the environment independent style could bring a considerable reduction in computational expense, however, the components which are statistically far from each other would be selected for mixture sharing. Selecting dissimilar components for sharing degrades the distinctiveness between the different environmental models, and would ultimately decrease the performance of feature compensation.

Here, we propose an environment dependent implementation for mixture sharing in an effort to take advantage of the environment transition model presented in Sect. 5 to minimize the performance loss in feature reconstruction. In the proposed scheme, the environmental models which are determined at the current speech (or session) by environment transition model are only considered for mixture sharing. Therefore, a different set of Gaussian components for mixture sharing is determined according to each environmental model \( e \). The procedure of selecting the similar components is presented as follows in pseudo code, where \( D_e \) is the set of distances between Gaussian components, and \( C_e \) is the resulting set of the shared components indices:

- **Step 0:** \( D_e = \{d_{1e}, d_{2e}, \ldots, d_{Ke}\}, C_e = \emptyset \)
  \[
  d_{ke} = \sum_{i,j}
  \end{equation}

- **Step 1:** \( k_{min} = \arg \min_{k} d_{ke} \in D_e \), \( D_e = D_e - \{d_{k_{min}}\} \)
- **Step 2:** \( C_e = C_e \cup \{k_{min}\} \), \( D_e = D_e - \{d_{k_{min}}\} \)
- **Step 3:** if \( N(C_e) = K_s \), then stop, else go back to **Step 1**.

In Eq. (6), \( c_{ij} \) is the \((e, i)\)th element of the connectivity matrix of the environment transition model in Eq. (5) and \( g_{lk} \) denotes the \( k \)th Gaussian component of the \( l \)th model among \( E \) number of environmental models. In the steps, \( d_{ke} \) is the sum of Kullback-Leibler (KL) distances of the \( k \)th Gaussian component across the environmental models which are connected with the \( e \)th environmental model and \( N(c) \) denotes the number of resulting shared elements.

For initialization, in **Step 0**, the sum of KL distances across the different environmental models is obtained per each Gaussian component index \( k \). Next, the most similar Gaussian index across the environmental models is selected in **Step 1**. Finally, the Gaussian search process is halted when the element number of the selected Gaussian set \( C_e \) reaches the desired number of sharing Gaussian components \( K_s \), which are now tagged as similar pdfs across the noisy speech models. When the \( e \)th environment is determined as the most likely observed condition in the previous speech, the parameters of the merged Gaussian components for sharing are computed as follows:

The likelihood functions which contain the Gaussian components included in set \( C_e \) are replaced by the merged Gaussian components,

\[
\begin{align*}
\mathbb{P}_{y|ke}^{[S]} &= \frac{1}{\sum_j c_{je}} \sum_j c_{je} \mu_{y,ijk}, k \in C_e \\
\mathbf{S}_{y|ke}^{[S]} &= \frac{1}{\sum_j c_{je}} \sum_j c_{je} (\mathbf{S}_{y,ijk} + (\mu_{y,ijk} - \mu_{y,ke}) (\mu_{y,ijk} - \mu_{y,ke})^T), k \in C_e.
\end{align*}
\]

The likelihood function of the Gaussian components included in set \( C_e \) is replaced by the merged Gaussian components,

\[
\begin{align*}
p(y|e, i, k) &= \begin{cases} p(y|\mu_{y,ke}, \Sigma_{y|ke}), & \text{if } k \in C_e \\ p(y|\mu_{y,ijk}, \Sigma_{y,ijk}), & \text{otherwise.} \end{cases}
\end{align*}
\]

where \( p(y|e, i, k) \) indicates the likelihood function of the \( k \)th component of the \( i \)th environmental model (i.e., \( G_i \) in Eq. (3)) when environmental model \( e \) is determined as the most likely observed condition in the previous speech.

The constant bias terms used for feature reconstruction in Eq. (4) are also shared if their indices are included in set \( C_e \),

\[
r_{i,ke} = \begin{cases} \mu_{y,ke}^{[S]} - \mu_{y,ki}, & \text{if } k \in C_e \\ \mu_{y,ijk} - \mu_{y,ki}, & \text{otherwise.} \end{cases}
\]

\[\text{where each index corresponds to each environment (i.e., noise condition, N1, ..., N7) and NX is considered as the 8th element. Each element in } C_{ETM}(i, j) \text{ can be written as } c_{ij}, \text{ and considered as prior knowledge on the } \text{7th environment for multiple environmental models in the current speech (or frame, session) when the } \text{6th environment is determined as the most likely observed condition in the previous speech.} \]

\[\text{The environment transition model is applied to the current input speech based on the type of environment most likely observed at the previous utterance, assuming that the current speech was produced in a continuing driving condition from the previous utterance. The most likely observed environment is determined by the maximum of the posterior probability of an environment } p(G_i|Y) \text{ over the input speech } Y \text{. Suppose that the N4 condition (highway driving, windows closed) was determined at the previous utterance. Only four noise conditions (N3, N4, N5, and NX) are considered for multiple environmental models by replacing the prior probabilities } P(G_i) \text{ in Eq. (3) with the normalized connectivity } c_{ij} / \sum_j c_{ij} \text{ of the environment transition model in Eq. (5). This case is expected to result in a reduced computational expense (= 4/9 amount) compared to the case of not employing the transition model (i.e., fully connected environmental models).} \]

\[\text{In } \text{Eq. (6), } c_{ij} \text{ is the } (e, i)\text{th element of the connectivity matrix of the environment transition model in Eq. (5) and } g_{lk} \text{ denotes the } k \text{th Gaussian component of the } l \text{th model among } E \text{ number of environmental models. In the steps, } d_{ke} \text{ is the sum of Kullback-Leibler (KL) distances of the } k \text{th Gaussian component across the environmental models which are connected with the } e \text{th environmental model and } N(c) \text{ denotes the number of resulting shared elements.} \]

\[\text{For initialization, in **Step 0**, the sum of KL distances across the different environmental models is obtained per each Gaussian component index } k \text{. Next, the most similar Gaussian index across the environmental models is selected in **Step 1**. Finally, the Gaussian search process is halted when the element number of the selected Gaussian set } C_e \text{ reaches the desired number of sharing Gaussian components } K_s \text{, which are now tagged as similar pdfs across the noisy speech models. When the } e \text{th environment is determined as the most likely observed condition in the previous speech, the parameters of the merged Gaussian components for sharing are computed as follows,} \]

\[\begin{align*}
\mathbb{P}_{y|ke}^{[S]} &= \frac{1}{\sum_j c_{je}} \sum_j c_{je} \mu_{y,ijk}, k \in C_e \\
\mathbf{S}_{y|ke}^{[S]} &= \frac{1}{\sum_j c_{je}} \sum_j c_{je} (\mathbf{S}_{y,ijk} + (\mu_{y,ijk} - \mu_{y,ke}) (\mu_{y,ijk} - \mu_{y,ke})^T), k \in C_e.
\end{align*}\]

\[\text{The likelihood functions which contain the Gaussian components included in set } C_e \text{ are replaced by the merged Gaussian components,} \]

\[p(y|e, i, k) = \begin{cases} p(y|\mu_{y,ke}, \Sigma_{y|ke}), & \text{if } k \in C_e \\ p(y|\mu_{y,ijk}, \Sigma_{y,ijk}), & \text{otherwise.} \end{cases}\]

\[\text{where } p(y|e, i, k) \text{ indicates the likelihood function of the } k \text{th component of the } i \text{th environmental model (i.e., } G_i \text{ in Eq. (3)) when environmental model } e \text{ is determined as the most likely observed condition in the previous speech.} \]

\[\text{The constant bias terms used for feature reconstruction in Eq. (4) are also shared if their indices are included in set } C_e, \]

\[r_{i,ke} = \begin{cases} \mu_{y,ke}^{[S]} - \mu_{y,ki}, & \text{if } k \in C_e \\ \mu_{y,ijk} - \mu_{y,ki}, & \text{otherwise.} \end{cases}\]
The computations over the $E \times K$ number of Gaussian likelihood functions is reduced to $K_s + E(K-K_s)$, leading to a computational reduction by as much as $(E-1)K_s$ via the presented sharing technique. If too many components are shared, performance degradation can result and therefore the number $K_s$ must be selected to balance computational savings versus system performance. Figure 4 illustrates the concept of the mixture sharing technique, where $C_e$ represents the resulting shared pdf when the environment $e$ is decided to be most likely observed in the previous speech.

7. Experimental Results

7.1 Baseline Performance

As test data for performance evaluation, the connected single digits portions from CU-Move corpus were selected. The test set consists of 949 utterances (length of 1 hour and 40 min) spoken by 20 different speakers (9 males and 11 females) in real-life in-vehicle conditions, which were collected in Minneapolis, Minnesota [11]. Speech samples were down-sampled to 8kHz and reflect an average 8.48 dB SNR obtained by the NIST STNR Speech Quality Assurance software [22].

The test set is an identical task to the Aurora2 evaluation framework so that Aurora2 evaluation toolkit was employed to evaluate system performance [23]. The speech samples are connected English-digits consisting of eleven words. Each word is represented by a continuous density HMM with 16-states and 3-mixtures per state. In addition to the digits, two silence models (i.e., normal silence and short pause) are used. The HMM parameters and the clean Gaussian mixture model for feature compensation methods were estimated using 8,840 clean speech training samples which are included in Aurora2.

The feature extraction algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed for the experiments [24]. An analysis window of 25 msec duration is used with a 10 msec skip for 8-kHz speech data. The computed magnitude spectrum is passed through a Mel-scaled filter-bank and 23 Mel-filter-bank outputs are transformed to 13 cepstral coefficients. The 0th cepstral coefficient was used instead of log energy, for the sake of convenience in model combination implementation. After extracting the 13th order cepstrum, the first and second order time derivatives are included during the decoding procedure (a total 39 dimensional feature vector).

Table 1 Performance of baseline and existing methods on CU-Move corpus.

<table>
<thead>
<tr>
<th>Method</th>
<th>WER (%)</th>
<th>Relative (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>70.02</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>60.10</td>
<td>14.17</td>
</tr>
<tr>
<td>SS+CMN</td>
<td>39.92</td>
<td>42.99</td>
</tr>
<tr>
<td>RATZ</td>
<td>39.26</td>
<td>43.93</td>
</tr>
<tr>
<td>AFE</td>
<td>31.45</td>
<td>55.08</td>
</tr>
<tr>
<td>VTS</td>
<td>48.31</td>
<td>31.01</td>
</tr>
<tr>
<td>VTS+SS+CMN</td>
<td>23.29</td>
<td>66.74</td>
</tr>
</tbody>
</table>

Here, 8 different types of noise samples (a total of 2 hours) were used for training noise models. As discussed in Sect. 5, noise data were collected during a route driving in city and rural area aimed at obtaining all types of actual driving conditions, which are recorded separately from the CU-Move corpus with an identical recording setup [18].

The performance of the baseline system (no compensation) is examined with comparison to several existing pre-processing algorithms in terms of environmental robustness for speech recognition. Spectral Subtraction (SS) and Cepstral Mean Normalization (CMN) were selected as conventional algorithms. They represent the most commonly used techniques for additive noise suppression and removal of channel distortion respectively. In spectral subtraction, the subtraction factor and flooring factor are set at 4.0 and 0.2 respectively, and background noise is estimated using the minimum statistics method with a time delay of approximately 250 msec [25]. For cepstral mean normalization, the average value of the cepstrum over the current input utterance was subtracted from each frame.

As one of the data-driven methods, RATZ [14], [19] was evaluated for performance comparison. Artificially generated noise-corrupted data using identical noise samples to the ones used for PCGMM in the next section were used for training the model for RATZ. The noise sample was randomly chosen from 8 types of noise described in Sect. 5 and added to the clean speech. The resulting training data for RATZ has an 8.30 dB SNR (i.e., NIST STNR) on average which is similar to the SNR of the test speech (8.48 dB). AFE (Advanced Front-End) algorithm developed by ETSI was also evaluated as one of the state-of-the-art methods, which contains an iterative Wiener filter andblind equalization [26]. We also evaluated another feature compensation method, VTS (Vector Taylor Series) for performance comparison where the noisy speech GMM is adaptively estimated using the EM algorithm over each test utterance [14]. Table 1 demonstrates performance of the baseline system and existing algorithms.

7.2 Evaluation of Basic PCGMM Methods

The performance of the PCGMM-based scheme was eval-
uated using identical conditions to the baseline test. The GMM of the clean speech for PCGMM was estimated using clean speech samples identical to those used for training the HMM. The clean speech model consists of 128 Gaussian components with diagonal covariance matrices. The noise model used for model combination has a single Gaussian model and its prior model was obtained by off-line training.

In Table 2, PCGMM indicates PCGMM-based feature compensation method using a single noise model which is trained off-line. The result of PCGMM is the average of the evaluation results when each different type of noise model (N1, N2, N7 or NX) is used for a single prior model. For PCGMMa, the mean of noise model is updated with the sample mean of silence of each test utterance to adapt the noise environment. Approximately 200 msec duration of silence is assumed to exist prior to the beginning of speech in every test utterance. This is a simulated case of employing an accurate silence detection algorithm. PCGMMa+SS+CMN denotes the PCGMMa method combined with Spectral Subtraction and CMN.

As presented in Table 2, the PCGMM-based feature compensation method is effective for in-vehicle conditions, and superior performance of the PCGMM method is demonstrated compared to spectral subtraction combined with CMN from Table 1. The results confirm that model combination used for estimation of the noisy speech GMM is effective for in-vehicle conditions, with Spectral Subtraction and CMN.

In Table 2, PCGMM indicates PCGMM-based feature compensation method using a single noise model which is trained off-line. The result of PCGMM is the average of the evaluation results when each different type of noise model (N1, N2, N7 or NX) is used for a single prior model. For PCGMMa, the mean of noise model is updated with the sample mean of silence of each test utterance to adapt the noise environment. Approximately 200 msec duration of silence is assumed to exist prior to the beginning of speech in every test utterance. This is a simulated case of employing an accurate silence detection algorithm. PCGMMa+SS+CMN denotes the PCGMMa method combined with Spectral Subtraction and CMN.

As presented in Table 2, the PCGMM-based feature compensation method is effective for in-vehicle conditions, and superior performance of the PCGMM method is demonstrated compared to spectral subtraction combined with CMN from Table 1. The results confirm that model combination used for estimation of the noisy speech GMM is effective in representing the noise corruption process. Relative improvement of 52.11% over baseline in WER was obtained through updating the mean of the noise model (PCGMMa), which is comparable or better to RATZ, AFE and VTS. PCGMMa+SS+CMN has a relative improvement 67.37% in WER which outperforms all other existing methods.

Using the identical setup, performance evaluation of the multiple environmental model schemes for PCGMM was also conducted. In the multi-model PCGMM method, 9 types of environmental models (N1, N2, ..., N7, NX, and clean condition) were used for model combination to generate noisy speech GMMs. As presented in Table 3, we see that PCGMM-based feature compensation schemes with multiple models are effective for in-vehicle conditions, with superior performance over existing conventional algorithms. PCGMM-based feature compensation with multiple models (mPCGMM) produces comparable performance to the single model adaptation approach (PCGMMa) from Table 2. This confirms that employing multiple models is very effective for compensating the feature adaptively under blind noise environments and changing noise types in every utterance. A significant improvement was obtained by combining the mPCGMM method with spectral subtraction and CMN.

### Table 2 Performance of PCGMM-based methods with single model.

<table>
<thead>
<tr>
<th>Method</th>
<th>WER (%)</th>
<th>Relative (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCGMM</td>
<td>42.39</td>
<td></td>
</tr>
<tr>
<td>PCGMMa</td>
<td>33.53</td>
<td></td>
</tr>
<tr>
<td>PCGMMa+SS+CMN</td>
<td>22.85</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3 Performance of PCGMM-based methods with multiple environmental models.

<table>
<thead>
<tr>
<th>Method</th>
<th>WER (%)</th>
<th>Relative (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mPCGMM</td>
<td>33.84</td>
<td></td>
</tr>
<tr>
<td>mPCGMM+SS+CMN</td>
<td>22.56</td>
<td></td>
</tr>
</tbody>
</table>

7.3 Evaluation of Proposed Schemes

Table 4 presents the performance evaluation of the multi-model PCGMM-based methods employing the environment transition model and mixture sharing technique described in Sect. 5 and Sect. 6 respectively. The test utterances are submitted to the speech recognizer in the same time-order as recorded in-vehicle. With the environment transition model shown in Fig. 3, for a particular speaker, it is determined which noise types are considered at the current utterance for multiple environmental models, based on the noise condition which has the highest score (a posteriori) at the previous utterance.

From Table 4, the multi-model PCGMM with the environment transition model (mPCGMM+ETM) demonstrates a comparable performance to the case of fully-connected environmental models (mPCGMM). In order to investigate the relationship between recognition performance and computational expense brought by employing the environment transition model, the average number of activated environmental models and resulting computational reduction are presented also in the last three columns in Table 4. The number of active models (# of Active Models) is the average number of environments considered at each frame by employing the environmental transition model. The computational reduction was calculated comparing to the fully-connected model which has 9 activated conditions. From the results, it was found that employing the environment transition model (ETM) is effective in significantly reducing the computational complexity by 30.89% while holding the original performance at comparable levels (i.e., 0.19% difference in relative improvement).

Table 4 also shows the performance when employing the mixture sharing technique. mPCGMM-EI32(64, 96)+ETM indicates the case of 32(64, 96) Gaussian components selected for the mixture sharing technique in the environment independent (EI) style which also employs the environment transition model. In EI mixture sharing, the most 32(64, 94) similar Gaussian components across all 9 environments are selected. The second to the last column (# of Gaussian) indicates the number of Gaussian components to be computed for the mPCGMM method, which has 1,152 (= 128 × 9) in the case of non-sharing. The mixture sharing case has Ks+8(128-Ks)+128 as the Gaussian components to be computed, because the clean condition is not included for mixture sharing.

From the results, as we increase the number of shared components, WER performance decreases. This is because increasing the number of shared components degrades the
distinctive property between the different environment models. We obtained considerable reduction in computational expenses, by employing mixture sharing. In the case of mPCGMM-EI32+ETM, a 48.09% computational reduction was obtained while showing only a 1.51% performance decrease. However, we see that mPCGMM-EI96+ETM suffers from a 7.74% loss in recognition performance while it obtains a significant advantage in computation (i.e., 73.56% reduction). The next experimental results show that this problem can be addressed by employing environmental dependent (ED) style of mixture sharing technique.

The mPCGMM-ED32[64, 96]+ETM system configuration in Table 4 indicates the ED implementation of the mixture sharing method. In this case, the connected environmental models in the environmental transition model are only considered for mixture sharing. For example, if N4 was the highest score of the environment at the previous speech (or session), only N3, N4, N5, and NX are considered for mixture sharing at the current speech. The results of the ED method show considerably different fashion in the performance from the EI cases discussed so far. Here, mPCGMM-ED96+ETM shows only a 0.83% decrease in recognition performance with a 73.10% computational reduction. This indicates that the ED style of mixture sharing is significantly more effective in maintaining the distinctive characteristics between the different environmental models while sharing a large number of Gaussian components. Such computational savings can be useful for limited CPU based speech applications running on PDAs or cellphone platforms.

Table 5 shows the performance results of the same experiment as in Table 4 combined with spectral subtraction and CMN. The results also indicate that the ED mixture sharing method is very effective in reducing the computational expense while maintaining recognition performance. Here, mPCGMM-ED96+ETM+SS+CMN shows only a 0.57% relative loss in recognition, while bringing a 67.96% computational reduction comparing to mPCGMM+SS+CMN which does not employ the environment transition model and mixture sharing technique. Comparing VTS+SS+CMN in Table 1, mPCGMM-ED96-ETM+SS+CMN also has an advantage in computational complexity, having slightly better performance by 0.33% WER. The reason is that VTS requires a number of iterations for EM estimation of the noise components, and it operates in the log-spectral domain which needs more coefficients compared to the cepstral domain. In our experiment, 23 log-spectral coefficients, 128-component GMM, and 5 iterations for EM were used for VTS algorithm. Which means that 640 (=5×128) numbers of Gaussian components must be computed for the 23-order coefficients. The proposed mPCGMM-ED96-ETM has 332 (=6.92×480) Gaussians for computation over the 13-order cepstral coefficients.

8. Conclusions

In this study, the PCGMM (Parallel Combined Gaussian
Mixture Model)-based feature compensation algorithm was evaluated on the CU-Move corpus, which contains a range of background noise observed in real-life in-vehicle conditions. The interpolation method of multiple environmental models was employed to address time-varying in-vehicle noisy conditions. To reduce the computational complexity due to multiple models, we employed an environment transition model, which determines a smaller set of environmental models for the PCGMM approach with multiple models. For more effective reduction in computational expenses, we proposed an environment dependent mixture sharing technique that takes advantage of our environment transition model. Similar Gaussian components only among the environmental model set determined by the environment transition model are selected for mixture sharing. Experimental results demonstrated that our feature compensation is effective in accomplishing reliable and efficient speech recognition in real-world in-vehicle environments. The mixture sharing scheme proposed in our study brought a significant computational reduction with only a slight loss in performance. This shows that the environment transition model presented in this paper is effective in maintaining the distinctive characteristic between the different environmental models while a large number of Gaussian components are selected for mixture sharing.

References


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

John H. L. Hansen 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 the Georgia Institute of Technology, Atlanta, 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 a Department Chairman of Electrical Engineering and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Brain and Behavioral 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, 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 39 (18 Ph.D., 21 M.S.) thesis candidates. He is author/coauthor of 255 journal and conference papers 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, 2007), and lead author of the report “The Impact of Speech Under ‘Stress’ on Military Speech Technology,” (NATO RTO-TR-10, 2000). Prof. Hansen is an IEEE Fellow, served as an IEEE Signal Processing Society Distinguished Lecturer for 2005/2006, is a member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee, and has served as Technical Advisor to the U.S. Delegate for NATO (IST/TG-01). He was an Associate Editor for the IEEE Trans. on Speech & Audio Proc. (1997–1999), an Associate Editor for IEEE Signal Proc. Letters (1998–2000), and an 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 Trans. on Speech & Audio Proc. 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 (2004–2010). He also organized and served as General Chair for ICSLP-2002, September 16–20, 2002, and will serve as Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.