FEATURE COMPENSATION EMPLOYING MODEL COMBINATION FOR ROBUST SPEECH RECOGNITION IN IN-VEHICLE ENVIRONMENT

Wooil Kim and John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim, John.Hansen}@utdallas.edu, http://crss.utdallas.edu

ABSTRACT

An effective feature compensation method is evaluated for reliable speech recognition in real-life in-vehicle environment. CU-Move corpus contains a range of speech and noise signals collected for a number of speakers under actual driving conditions. PCGMM-based feature compensation, considered in this paper, utilizes parallel model combination to generate noise-corrupted speech model by combining clean speech and the noise model. In order to address unknown time-varying background noise, an interpolation method of multiple environmental models is employed. To alleviate computational expenses due to multiple models, applying a noise transition model is proposed, which is motivated from *Noise Language Model* used in *Environmental Sniffing*. The PCGMM method and proposed scheme are evaluated on the connected single digits portion of the CU-Move database using Aurora2 evaluation toolkit. Experimental results indicate that our feature compensation method is effective for improving speech recognition in real-life in-vehicle conditions. Here, 26.78% of the computational reduction was obtained by employing the noise transition model with only slight change in recognition performance.

1. INTRODUCTION

Acoustic difference between training environments and conditions where actual speech recognition systems operate is one of the primary factors that degrade speech recognition accuracy, and the presence of background noise is one major factor. This is especially true for in-vehicle speech systems which face the problem of robust speech recognition in order to address a range of severe changing background noise conditions.

This paper investigates the performance of our feature compensation scheme in a real-life in-vehicle environment, with the goal of achieving low complexity in computation. 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 [1]. The corpus has been used for research in multi-sensor array processing for noise suppression and speech recognition in cars [2]. Therefore, performance evaluation on CU-Move database 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 [3] 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. Independent access to the noise model makes its adaptation in the non-speech interval possible. The interpolation method employing multiple environmental noise models was also developed to address unknown or time-varying noise conditions [4]. In order to reduce the computational expense due to use of multiple models, we propose to employ a noise transition model for the multi-model approach, which is motivated from *Noise Language Model* in our previous work [5].

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

2. CU-MOVE CORPUS

The CU-Move project [1] is 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. The database and noise conditions are discussed in detail in [1]. 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 10 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

The PCGMM-based feature compensation method is based on the speech model. The distribution of the clean speech feature x in the cepstral domain is represented with a Gaussian Mixture Model (GMM) consisting of K components as follows,

$$p(x) = \sum_{k=1}^{K} \omega_k N(x; \mu_{x,k}, \Sigma_{x,k}) \quad . \tag{1}$$

It is assumed that noisy environment degrades by moving the means and covariance matrices of the clean speech model, and the distribution of the noisy speech y can be also expressed as,

$$p(y) = \sum_{k=1}^{K} \omega_k N(y; \mu_{y,k}, \Sigma_{y,k}) \cdot \quad (2)$$

In PCGMM-based method, the parameters of the noise-corrupted speech model $\mu_{y,k}$ and $\Sigma_{y,k}$ are obtained through parallel model combination (PMC) procedure using clean speech and noise models independently [3]. It is also assumed that there is a constant bias transformation of the mean parameters of the clean speech model in the cepstral domain under the additive noisy environment, which is the assumption generally taken by other data-driven methods as follows,

$$\mu_{y,k} = \mu_{x,k} + r_k \quad (3)$$

where the bias term r_k is used for reconstruction of the speech features. The MMSE equation for reconstruction of the clean speech is approximated with Eq.(4) in a manner similar to [6].

$$\hat{x}_{MMSE} = \int_X x p(x \mid y) dx \cong y - \sum_{k=1}^{K} r_k \, p(k \mid y) \quad (4)$$

The posterior probability $p(k \mid y)$ can be calculated using the parameters of the noisy speech GMM $\{\omega_k, \mu_{y,k}, \Sigma_{y,k}\}$. Fig 1 presents the resulting block diagram of the PCGMM-based approach as described here.

At this point, the distinguishing properties of the PCGMM-based method are considered, and compared with prior techniques. First, our method does not require an additional training procedure using a noise-corrupted speech database. After obtaining the estimated noise model from the available noise samples, the distribution model of the noise-corrupted speech can be generated via the model combination procedure. This results in a compensation method without the need of prior training data as seen in existing data-driven methods.

In the PCGMM method, estimation of the GMMs for clean speech, noise, and noisy speech as well as the reconstruction procedure are accomplished all in the cepstral domain. The number of cepstral coefficients is generally smaller than for log-spectral coefficients, therefore, our method has the explicit advantage of a lower dimensional space (e.g., reduced computation). In particular, the cepstral coefficients are less correlated with each other compared to the same coefficients in the log-spectral domain, therefore it is reasonable to employ diagonal covariance matrices for the GMMs in representing the models. The movement from a full covariance matrix needed for the log-spectral domain to a diagonal covariance matrix in the cepstral domain has a major reduction in both computational costs and input data requirements for more accurate model estimation.

4. PCGMM-BASED METHOD EMPLOYING MULTIPLE ENVIRNMENTAL MODELS

In the PCGMM-based method, model adaptation can be applied in order to address the time-varying background noise. In such a framework, the noise model is updated during silence periods via adaptation followed by combination of models, which again more accurately 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 domain. Therefore, applying a model adaptation technique for the noise model may not be appropriate for small resource systems such as PDAs, navigation devices and other mobile systems. In this section, we consider the PCGMM-based method that employs a combination of multiple environmental models for low resource based ASR applications.

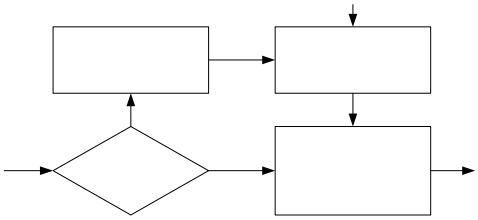

Fig. 1. Block diagram of PCGMM-based feature compensation method.

Utilizing multiple models estimated off-line can 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 [4]. Given the incoming noisy speech feature vectors $Y_t = [y_1, y_2, \ldots, y_t]^T$, the sequential posterior probability of a specific environment GMM G_i among E models over the input speech feature Y_t can be re-written as,

$$p(G_i \mid Y_t) = \frac{P(G_i)\, p(Y_{t-1} \mid G_i)\, p(y_t \mid G_i)}{\sum_{e=1}^{E} P(G_e)\, p(Y_{t-1} \mid G_e)\, p(y_t \mid G_e)} \quad (5)$$

where $p(Y_{t-1} \mid G_i) = \prod_{\tau=1}^{t-1} p(y_\tau \mid G_i)$ and $P(G_i)$ is a prior probability of each environment i represented as a GMM. Based on Eq.(4), the clean feature at frame t is reconstructed using the interpolated compensation terms as follows,

$$\hat{x}_{t,MMSE} \cong y_t - \sum_{e=1}^{E} p(G_e \mid Y_t) \sum_{k=1}^{K} r_{e,k} \, p(k \mid G_e, y_t) \quad (6)$$

where $r_{e,k}$ is a constant bias term from the k th Gaussian component of the e th environment model and $p(k \mid G_e, y_t)$ is the posterior probability for environment G_e.

When the background noise is 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. [5]), the multiple-model method is more effective than adaptation techniques or online estimation of noise components in terms of computational complexity. In time varying scenarios, it is also possible to employ *Environmental Sniffing* to detect, track, and characterize the noise types [5]. If a clean mixture model is considered as one of the multiple models, the performance of the recognition system can be maintained under high Signal-to-Noise Ratio (SNR) conditions.

5. NOISE TRANSITION MODEL

The amount of computation for model-based feature compensation depends primarily on the number of Gaussian components to be computed. The computational expense increases in proportion to the number of multiple models employed for the model interpolation method described in Sec.4. However, more accurate modeling for noisy conditions requires a larger number of GMMs with sufficient sized pdfs. Now, we describe a noise transition model employed in an effort to reduce the computational complexity.

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 noise models for PCGMM-based feature compensation. This can reduce the computational expenses. In the *Environmental Sniffing* scheme, a *Noise Language Model* was employed to decode the most likely sequence of noise types [5]. Here, the noise transition model is motivated from the noise language model. In order to build the noise language model, in-vehicle acoustic data (i.e., in a Blazer SUV) was collected during a 17-mile route driving which contains samples of all driving conditions expected for use in city and rural areas and then the primary noise conditions were identified as follows:

(1) N1: idle noise, no movement, windows closed
(2) N2: city driving without traffic, windows closed
(3) N3: city driving with traffic, windows closed
(4) N4: highway driving, windows closed
(5) N5: highway driving, windows 2 inches open
(6) N6: highway driving, windows half-way down
(7) N7: windows 2 inches open in city traffic
(8) NX: others

A bigram type of noise language model was constructed using CMU-Cambridge Statistical Language Modeling (SLM) Toolkit. In this study, the connectivity among the noise conditions was employed for the transition model not considering transition probabilities. Fig 2 shows the noise transition model considered in this paper.

The transition model is applied to the current speech input based on the type of noise observed at the previous utterance when the current speech was produced in a continuing driving condition from the previous utterance. Suppose that the N4 condition (highway driving, windows closed) was determined at the previous utterance by the accumulated posterior probability, only four type conditions (N3, N4, N5, and NX) are considered for multiple environmental models by setting the prior probabilities $P(G_i)$ in Eq.(6) of the other four conditions (N1, N2, N6, and N7) to zero according to the connectivity of noise transition model as shown in Fig.2. In this case, we expect to have a reduced computational expense comparing to case of not employing a noise transition model (i.e., fully connected noise models).

6. EXPERIMENTAL RESULTS

As test data for performance evaluation, connected single digits portions from CU-Move corpus were selected. They have an identical task to the Aurora2 evaluation framework so that Aurora2 evaluation toolkit was used to evaluate system performance [7]. The task is connected English-digits consisting of eleven words. Each whole 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 feature extraction algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed for the experiments [8]. 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 of 39 dimensional feature vector).

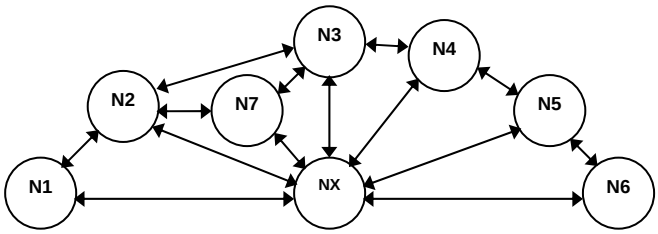

Fig. 2. Noise transition model.

Table 1. Performance of baseline system and existing methods on CU-Move Corpus. (WER, %)

Baseline	64.77
SS	55.13
SS+CMN	40.39
AFE	31.73
VTS	33.52
VTS+SS+CMN	26.33

The HMM parameters were estimated using 8,840 clean speech training samples included in Aurora2 and performance was evaluated on the selected test set of CU-Move corpus. The test set consists of 464 utterances (length of 50.0min) spoken by 10 different speakers (5 males and 5 females) in real-life in-vehicle conditions, which were collected in Minneapolis, Minnesota [1]. Data was down-sampled to 8kHz and reflected a 9.50 dB SNR on average which was obtained by NIST Speech Quality Assurance software [9]. We used 8 different types of noise samples (total amount of 2 hours) to train noise models, which were discussed in Sec. 5.

The performance of the baseline system (no compensation) is examined with comparison to several existing preprocessing 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 250msec. For cepstral mean normalization, the average value of the cepstrum over the current input utterance was subtracted from each frame. AFE (Advanced Front-End) algorithm developed by ETSI was also evaluated as one of state-of-the-art methods, which contains an iterative Wiener filter and cepstral histogram equalization [10]. We also evaluated another feature compensation method, VTS (Vector Taylor Series) algorithm for performance comparison where the noisy speech GMM is adaptively estimated using the EM algorithm over each test utterance [6]. Table 1 demonstrates performance of the baseline system and existing algorithms.

The performance of the PCGMM-based scheme was evaluated 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

Table 2. Performance of PCGMM-based methods.

	WER (%)	Relative (%)
PCGMM	62.31	3.80
PCGMMm	33.64	48.06
PCGMMm+SS+CMN	25.38	60.82

Table 3. Performance of multi-model PCGMM-based methods.

	WER (%)	Relative (%)
IM-PCGMM	34.08	47.38
IM-PCGMM+SS+CMN	25.83	60.12

offline-training. For the prior noise model for single model PCGMM, the noise signals from the type NX were used, which has connections between all other noise types. For comparison, we examined the performance in the following combinations:

(1) **PCGMM**: PGCMM-based feature compensation method using model combination of clean speech model and prior noise model trained off-line.

(2) **PCGMMm**: the mean of noise model is updated with the sample mean of silence of each test utterance for PCGMM method. Approximately 200msec duration of the silence is assumed to exist prior to the beginning of speech in every test utterance.

(3) **PCGMMm+SS+CMN**: PCGMMm 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 in Table 1. The results prove that the model combination used for the estimation of noisy speech GMM is effective in representing the noise corruption process. Relative improvement of 48.06% over baseline in WER was obtained through updating the mean of the noise model (PCGMMm), which is better or comparable to AFE and single VTS. PCGMMm method combined with spectral subtraction and CMN has a relative improvement of 60.82% in WER and it outperforms all other existing methods.

Using the same setup, performance evaluation of the multi-model schemes for PCGMM was also conducted. In the interpolation of multi-model PCGMM method, 9 types of noise models (N1, N2, … , N7, and NX including 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 the interpolation method of multiple models are effective for the in-vehicle conditions, with superior performance over existing conventional algorithms. The PCGMM-based feature compensation with interpolated models (IM-PCGMM) presents comparable performance to the single adapted model approach (PCGMMm) which is shown in Table 2. This proves that interpolation of multiple models is very effective for compensating the feature adaptively under blind noisy environments and changing noise types in every utterance. A significant improvement was obtained by combining the IM-PCGMM method with spectral subtraction and CMN.

Tables 4 and 5 present the performance of the IM-PCGMM-based method employing the noise transition model described in Sec.5. The test utterances are submitted to the speech recognizer in the same time-order as recorded in-vehicle. With the noise transition model, for a particular speaker, it is determined which noise types are considered for 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 IM-PCGMM with the noise transition model demonstrates a comparable performance compared to the case of fully-connected noise model in Table 3. In order to investigate the relationship between performance and computational expense brought by employing noise transition model, the average number of activated noise models and resulting computational reduction are presented in Table 5. The computational reduction was calculated comparing to the fully connected noise model which has 9 numbers of activated conditions. From the results, it was found that employing noise transition model is useful for reducing the computational complexities while holding the original performance at comparable levels.

Table 4. Performance of multi-model PCGMM-based methods with Noise Transition Model.

	WER (%)	Relative (%)
IM-PCGMM	33.86	47.72
IM-PCGMM+SS+CMN	25.98	59.89

Table 5. Computational reduction of multi-model PCGMM-based methods by employing Noise Transition Model.

	# of Activated Noise Models	Computational Reduction (%)
IM-PCGMM	6.22	30.89
IM-PCGMM+SS+CMN	6.59	26.78

7. CONCLUSIONS

In this paper, we evaluated the PCGMM-based feature compensation method on the CU-Move corpus which contains a range of background noise observed in real-life in-vehicle conditions. To reduce the computational complexity, employing a noise transition model was proposed for a multiple model approach. Experimental results demonstrate that our feature compensation is effective in accomplishing reliable and efficient speech recognition in actual in-vehicle environments.

REFERENCES

[1] J.H.L. Hansen, X. Zhang, M. Akbacak, U. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul, "CU-Move: Advanced in-vehicle speech systems for route navigation," *DSP for in-vehicle and mobile systems*, Springer-Verlag, 2004.

[2] , X. Zhang and J.H.L. Hansen, "CSA-BF: a constrained switched adaptive beamformer for speech enhancement and recognition in real car environments," *IEEE Trans. on Speech and Audio Proc.*, 11(6), pp.733-745, 2003.

[3] W. Kim, S. Ahn, and H. Ko, "Feature Compensation Scheme Based on Parallel Combined Mixture Model," *Eurospeech2003*, pp.667-680, 2003.

[4] W. Kim, O. Kwon, and H. Ko, "PCMM-based Feature Compensation Scheme Using Model Interpolation and Mixture Sharing," *ICASSP2004*, pp.989-992, 2004.

[5] M. Akbacak and J.H.L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems," *IEEE Trans. on Audio, Speech and Language Proc.*, 15(2), pp.465-477, 2007.

[6] P. J. Moreno, B. Raj, and R. M. Stern, "Data-Driven Environmental Compensation for Speech Recognition: A Unified Approach," *Speech Communication*, 24: 267-85, 1998.

[7] H. G. Hirsch & D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW ASR2000*, Sep. 2000.

[8] ETSI standard doc., "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," *ETSI ES 201 108 v1.1.2 (2000-04)*, 2000.

[9] NIST SPeech Quality Assurance (SPQA) package version 2.3, http://www.nist.gov/speech

[10] ETSI standard doc., "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," *ETSI ES 202 050 v1.1.1 (2002-10)*, 2002.