AN ADVANCED FEATURE COMPENSATION METHOD EMPLOYING ACOUSTIC MODEL WITH PHONETICALLY CONSTRAINED STRUCTURE

Wooil Kim[1] *and John H. L. Hansen*[2]

[1]School of Computer Science and Engineering, Incheon National University, Incheon, Korea
[2]Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, Texas, USA
wikim@incheon.ac.kr, John.Hansen@utdallas.edu

ABSTRACT

This study proposes an effective model-based feature compensation method for robust speech recognition in background noise conditions. In the proposed scheme, an acoustic model with a phonetically constrained structure is employed for the Parallel Combined Gaussian Mixture Model (PCGMM [1]) based feature compensation method. The structure of the acoustic model includes a collection of context independent phone models. A phonetically constrained prior probability is formulated by integrating transition probability of phone models into the reconstruction procedure. Experimental results show that the PCGMM-based feature compensation employing the proposed phonetically constrained structure of acoustic model consistently outperforms the case of employing the conventional Gaussian mixture model. This demonstrates that the proposed configuration of the acoustic model is effective at improving the intelligibility of the speech reconstructed by the feature compensation method for speech recognition under diverse background noise conditions.

Index Terms— feature compensation, PCGMM, acoustic model, phonetically constrained structure, robust speech recognition

1. INTRODUCTION

The presence of background noise generates acoustic mismatch between training and operating conditions in actual speech recognition systems, which is one primary factor severely degrading recognition performance. To minimize this mismatch, extensive research has been conducted in recent decades, which includes many types of speech/feature enhancement methods such as Spectral Subtraction, Cepstral Mean Normalization (CMN), and a variety of feature compensation schemes. 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) [2]-[4].

This study focuses on a feature compensation method employing an acoustic model for speech feature space which is usually estimated as a Gaussian Mixture Model (GMM). Multivariate Gaussian-Based Cepstral Normalization (RATZ), Vector Taylor Series (VTS) [5][6], and Stereo-based Piecewise Linear Compensation for Environments (SPLICE) [7] algorithms can be classified into this GMM-based feature compensation family, and our previous work presented in [1] is also based on the GMM for the acoustic model.

In this paper, we propose to improve the GMM-based feature compensation method by leveraging phonetic knowledge. The proposed method utilizes a collection of phone models instead of the general speech GMM (i.e., Universal Background Model (UBM) used for speaker recognition technique). It also integrates transition probability of the phone models which defines statistical connectivity among the phone models. A *phonetically constrained prior probability* of each phone model is calculated for each speech frame in the proposed algorithm. We believe that such a constrained structure of the acoustic model could represent phonetic dynamics of speech utterance and it would be more effective at improving the intelligibility of the reconstructed speech, compared to the conventional GMM-based method. Our previously proposed Parallel Combined GMM (PCGMM) feature compensation scheme is employed as a baseline framework in this work [1].

This paper is organized as follows. We first review the PCGMM-based feature compensation method as a framework for this study in Sec. 2. Sec. 3 presents the proposed approach employing a phonetically constrained acoustic model. Representative experimental procedures and their results are presented and discussed in Sec. 4. Finally, in Sec. 5 we summarize the main conclusions of our work.

2. PCGMM-BASED FEATURE COMPENSATION

In the PCGMM-based method [1], the parameters of the noise-corrupted speech model $\{\omega_k, \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\}$ are obtained

This work was supported by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

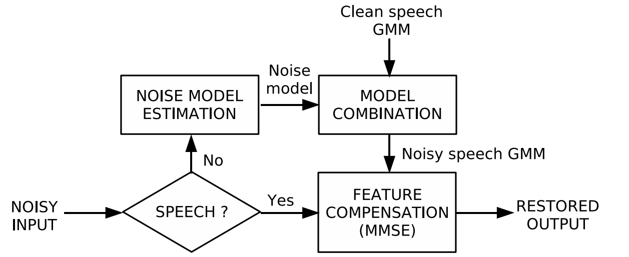


Fig. 1. *Block diagram of the PCGMM-based feature compensation method [1].*

through a model combination procedure using clean speech and noise models independently [4]. The clean speech model $\{\omega_k, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}\}$ consists of K Gaussian components, and the noise model is estimated with a single Gaussian pdf $\{\boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\}$ both in the cepstral domain. The noise-corrupted speech model is obtained as,

$$\{\omega_k, \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\} = \mathcal{F}[\{\omega_k, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}\}, \{\boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\}], \quad (1)$$

where $\mathcal{F}[\cdot]$ denotes a function representing the model combination, and the same weight parameter ω_k is just used as in the clean speech model.

A constant bias transformation of the mean parameters of the clean speech model is assumed in the cepstral domain under the additive noise environment, which is the assumption generally taken by other data-driven methods [5] as follows,

$$\boldsymbol{\mu}_{\mathbf{y},k} = \boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{r}_k, \quad (2)$$

where the bias term $\mathbf{r}_k$ is used for reconstruction of the input speech. The bias term is estimated by Eq. (2), once the mean parameters of the clean speech model and corresponding noise-corrupted speech model are obtained. The Minimum Mean Squared Error (MMSE) estimation equation for reconstruction of the clean speech is approximated as follows [5][1],

$$\hat{\mathbf{x}}_{MMSE} = \int_{\mathcal{X}} \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x} \cong \mathbf{y} - \sum_{k=1}^{K} \mathbf{r}_k \; p(k|\mathbf{y}). \quad (3)$$

The posterior probability $p(k|\mathbf{y})$ is given by,

$$p(k|\mathbf{y}) = \frac{\omega_k p(\mathbf{y}|k)}{\sum_{k=1}^{K} \omega_k p(\mathbf{y}|k)}, \quad (4)$$

where $p(\mathbf{y}|k) = p(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k})$. Fig. 1 presents the block diagram of the PCGMM-based feature compensation as presented in this section.

3. PCGMM SCHEME EMPLOYING ACOUSTIC MODEL WITH PHONETICALLY CONSTRAINED STRUCTURE

As a first step, context independent (CI) phone models are obtained from a clean speech training database. The ith phone model is represented as a GMM (i.e., an HMM with a single state) which consists of K Gaussian components as follows:

$$p(\mathbf{x}|i) = \sum_{k=1}^{K} \omega_{i,k} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x},i,k}, \boldsymbol{\Sigma}_{\mathbf{x},i,k}). \quad (5)$$

From the same training database, a prior probability of each model $p(i)$ and a model transition probability $p(i|j)$ are also obtained. The model transition probability $p(i|j)$ represents a probability of transition from a phone model j at a previous time frame $t-1$, to a model i at the current frame t.

Using the model combination procedure in the PCGMM method, noise-corrupted phone models are obtained for all corresponding clean phone models as follows:

$$p(\mathbf{y}|i) = \sum_{k=1}^{K} \omega_{i,k} \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{y},i,k}, \boldsymbol{\Sigma}_{\mathbf{y},i,k}), \quad (6)$$

where $\{\boldsymbol{\mu}_{\mathbf{y},i,k}, \boldsymbol{\Sigma}_{\mathbf{y},i,k}\}$ are generated by combining the clean model $\{\boldsymbol{\mu}_{\mathbf{x},i,k}, \boldsymbol{\Sigma}_{\mathbf{x},i,k}\}$ with a noise model. If all phone models are fully connected with an equal model transition probability, Eq. (3) can be formulated as,

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} - \sum_{i=1}^{I} p(i) \sum_{k=1}^{K} \mathbf{r}_{i,k} \; p(i,k|\mathbf{y}), \quad (7)$$

where I denotes the number of phone models. Here, the posterior probability $p(i,k|\mathbf{y})$ is given by,

$$p(i,k|\mathbf{y}) = \frac{\omega_{i,k} p(\mathbf{y}|i,k)}{\sum_{k=1}^{K} \omega_{i,k} p(\mathbf{y}|i,k)}, \quad (8)$$

where $p(\mathbf{y}|i,k) = p(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y},i,k}, \boldsymbol{\Sigma}_{\mathbf{y},i,k})$.

In the proposed scheme, we employ an acoustic model which is phonetically constrained by integrating the model transition probability $p(i|j)$. Here, we introduce $q_t(i)$ which is formulated by,

$$q_t(i) = p(i) \sum_{j=1}^{I} p(j) p(\mathbf{y}_{t-1}|j) p(i|j). \quad (9)$$

Fig. 2 illustrates the calculation of $q_t(i)$. The phonetically constrained prior probability $p_t^c(i)$ at time t is calculated by normalizing $q_t(i)$ as follows:

$$p_t^c(i) = \frac{q_t(i)}{\sum_{i=1}^{I} q_t(i)}. \quad (10)$$

Finally, the reconstruction equation is reformulated by replacing $p(i)$ with the proposed phonetically constrained prior probability $p_t^c(i)$:

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} - \sum_{i=1}^{I} p_t^c(i) \sum_{k=1}^{K} \mathbf{r}_{i,k} \; p(i,k|\mathbf{y}). \quad (11)$$

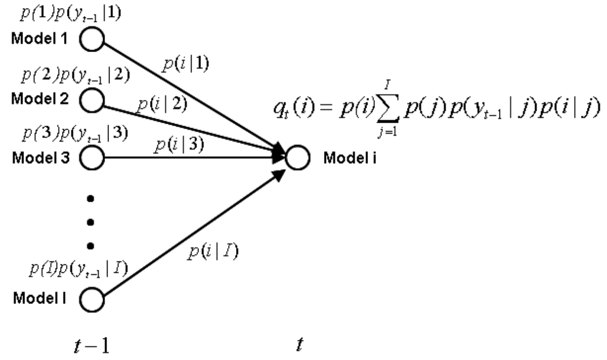
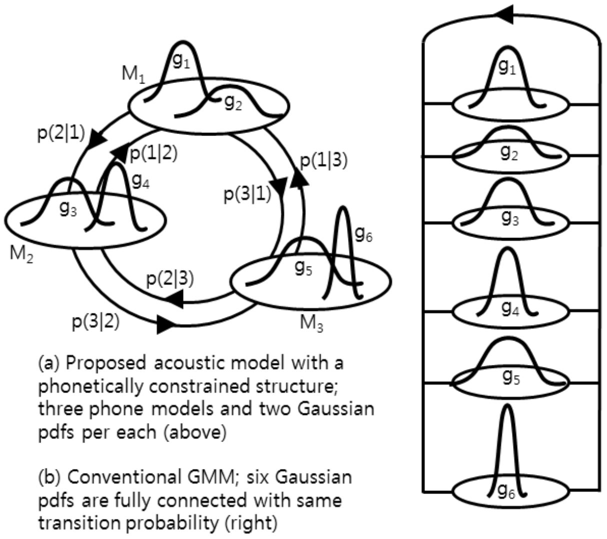


Fig. 2. *Calculation of $q_t(i)$ for the phonetically constrained prior probability.*



(a) Proposed acoustic model with a phonetically constrained structure; three phone models and two Gaussian pdfs per each (above)

(b) Conventional GMM; six Gaussian pdfs are fully connected with same transition probability (right)

Fig. 3. *Phonetically constrained structure of acoustic model.*

Fig. 3 illustrates an example structure of the acoustic model presented in this section. Here it is assumed that three phone models and two Gaussian pdf per a model. The conventional GMM with six Gaussian components is compared. We believe that this phonetically constrained structure of the acoustic model for feature compensation will be more effective at improving intelligibility of reconstructed speech.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup and Baseline Performance

The TIMIT speech corpus was used for performance evaluation of the proposed method. A total of 4.1 hours of speech (462 speakers, 4,620 utterances) were used for training, and 1.5 hours of data (168 speakers, 1,680 utterances) were used for test. The training and the test sets do not overlap each other in speakers or sentences. The data was down-sampled to 8kHz, so that each speech sample contains 4kHz 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 car noise and speech babble audio samples. The car noise and speech babble samples were obtained from NOISEX92 [8]. Each test set consists of 1,680 utterances at three different SNRs: 5, 10, and 15 dB.

Table 1. Recognition performance of baseline system and conventional methods for car noise and speech babble conditions (WER, %).

Car Noise	5 dB	10 dB	15 dB	Avg.
No Processing	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
ETSI AFE	48.20	29.88	20.24	32.77
Speech Babble	5 dB	10 dB	15 dB	Avg.
No Processing	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
ETSI AFE	50.68	30.72	19.89	33.76

We employed SPHINX3 [9] as the HMM-based speech recognizer to obtain recognition accuracy in background noise conditions. Each HMM represents a tri-phone which consists of 3 states with an 8-component GMM per state, which is tied with 1138 states. The task has 6233 vocabulary words, and the trigram language model is adapted on the TIMIT database using a Broadcast News language model as an initial model. A conventional MFCC (Mel-Frequency Cepstral Coefficient) feature front-end is employed in the experiment, which was suggested by the European Telecommunication Standards Institute (ETSI) [10]. An analysis window of 25msec is used with a 10msec 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 recognition system has 8.05 % Word Error Rate (WER) for clean speech conditions.

Performance of the baseline system (no processing) was examined with comparison to several existing pre-processing algorithms in terms of speech recognition performance. Spectral Subtraction (SS) [11] combined with Cepstral Mean Normalization (CMN) was selected as one of the conventional algorithms. These represent some of the most commonly used techniques for additive noise suppression and removal of channel distortion respectively. We also evaluated the Vector Taylor Series (VTS) algorithm for performance comparison [5]. 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 [12]. Table 1 demonstrates speech recognition performance (WER) of the baseline system and conventional algorithms on car noise and speech babble conditions. It shows that the ETSI AFE provides the greatest improvement in WER.

Table 2. Performance comparison of the PCGMM method employing different types of acoustic models (WER, %).

Car Noise	5 dB	10 dB	15 dB	Avg.
GMM with $K = 368$	52.87	31.33	17.36	33.85
Equal Transition	54.87	32.74	17.40	35.00
Phone Constrained	**52.67**	**30.47**	**17.34**	**33.49**
(Relative Improve)	**(+0.38)**	**(+2.74)**	**(+0.12)**	**(+1.08)**
Speech Babble	5 dB	10 dB	15 dB	Avg.
GMM with $K = 368$	51.99	26.84	15.94	31.59
Equal Transition	51.46	26.68	15.70	31.28
Phone Constrained	**50.74**	**25.95**	**14.79**	**30.49**
(Relative Improve)	**(+2.40)**	**(+3.32)**	**(+7.21)**	**(+4.31)**

4.2. Performance Evaluation of PCGMM Employing the Proposed Phonetically Constrained Structure of Acoustic Model

Next, Table 2 shows performance comparisons of different acoustic models which are employed by the PCGMM-based feature compensation method. For the proposed method, we obtained context independent (CI) phone models by training over the clean speech database. The CI phone models consists of 46 phones including a silence model, and each model is represented as a Gaussian mixture model (i.e., an HMM with a single state) which consists of 8 components, resulting in 368 Gaussian components in total. For a performance comparison, the baseline system employs a conventional GMM for the PCGMM method, which consists of 368 Gaussian components (i.e., "GMM with $K = 368$" in Table 2). We also evaluated the performance of the PCGMM method employing the identical CI phone models which are fully connected with an equal transition probability (i.e., "Equal Transition"). The scheme employing the "Equal Transition" phone models is implemented by Eq. (7).

By employing the phonetically constrained acoustic model for the PCGMM method, we obtained consistently improved performance compared to both the baseline (i.e., "GMM with $K = 368$") and equal transition model systems. We obtained +1.08% and +4.31% average relative improvements in WER compared to the baseline PCGMM system for car noise and speech babble conditions respectively. These results demonstrate that the proposed configuration of the acoustic model is effective at improving the intelligibility of the speech reconstructed by the feature compensation method for speech recognition under diverse background noise conditions.

5. RELATION TO PRIOR WORK

The existing model-based feature compensation methods mostly employ Gaussian mixture mode (GMM) as their acoustic model [5][6][7][1]. The proposed work in this paper employs an acoustic model with a phonetically constrained structure. The structure of the acoustic model consists of a collection of phone models. A phonetically constrained prior probability is formulated by integrating transition probability of phone models into the reconstruction procedure.

6. CONCLUSIONS

This study has proposed an effective model-based feature compensation method for robust speech recognition in background noise conditions. In the proposed scheme, an acoustic model with a phonetically constrained structure was formulated for the PCGMM-based feature compensation method. Context independent phone models and model transition probability parameters were obtained from a training database. The model transition probability was integrated into the reconstruction formulation for the feature compensation method. Experimental results demonstrated that the proposed configuration of the acoustic model is effective at improving the intelligibility of the speech reconstructed by the feature compensation method for speech recognition under background noise conditions.

7. REFERENCES

[1] W. Kim and J.H.L. Hansen, "Feature Compensation in the Cepstral Domain Employing Model Combination," *Speech Comm.*, 51(2), pp.83-96, 2009.

[2] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Proc.*, vol.2, no.2, pp.291-298, 1994.

[3] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," *Computer Speech and Language*, 9, pp.171-185, 1995.

[4] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on Speech and Audio Proc.*, vol.4, no.5, pp.352-359, 1996.

[5] P.J. Moreno, *Speech recognition in noisy environments*, Ph.D. Thesis. Carnegie Mellon University, 1996.

[6] P.J. Moreno, B. Raj, and R.M. Stern, "Data-driven Environmental Compensation for Speech Recognition: A Unified Approach," *Speech Communication*, 24(4), pp.267-285, 1998.

[7] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora 2 Database," *Eurospeech-2001*, pp. 217-220, 2001.

[8] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," *Tech. Rep., Speech Research Unit, Defense Research Agency*, Malvern, UK, 1992 (Available from NOISEX-92 CD-ROMS).

[9] http://cmusphinx.sourceforge.net

[10] *ETSI standard document*, ETSI ES 201 108 v1.1.2 (2000-04), 2000.

[11] R. Martin, "Spectral Subtraction based on Minimum Statistics," . *EUSIPCO-94*, pp. 1182-1185, 1994.

[12] *ETSI standard document* ETSI ES 202 050 v1.1.1 (2002-10), 2002.