Gaussian Map based Acoustic Model Adaptation Using Untranscribed Data for Speech Recognition in Severely Adverse Environments

Wooil Kim and John H. L. Hansen

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

Abstract

This study proposes an acoustic model adaptation scheme to improve speech recognition in severely adverse environments utilizing untranscribed data. In the proposed method, a clean GMM is estimated from clean training data, and a noise-corrupted GMM is obtained by MAP adaptation over the adaptation data. The Gaussian component of the adapted HMMs is obtained using the transform of the most similar Gaussian component of the GMM. The proposed mixture-selective model adaptation method is evaluated using an LDC corpus which represents severely adverse communication channel environments. The experimental results show the proposed adaptation method is comparable or improves performance compared to conventional MLLR adaptation. The proposed method is also effective at improving speech recognition using independent adaptation data sets. Performance results demonstrate that the proposed adaptation method is significantly more effective at improving speech recognition in severely noise conditions, where transcribed data is unavailable and baseline ASR fails to accurately transcribe the adaptation data due to acoustic condition mismatch.

Index Terms: model adaptation, untranscribed data, Gaussian mapping, adverse environments, robust speech recognition.

1. Introduction

Mismatch between training and operating conditions for any actual speech recognition system is one of the primary factors that severely degrades recognition accuracy. Background noise, microphone mismatch, communication channel, and speaker variability are major sources of such mismatch. Recently, as mobile devices such as smart phones have become popular, speech recognition technology via mobile systems is becoming more challenging, since a range of background noise and time-varying channel effects make recognition more difficult. This study focuses on an acoustic model adaptation scheme for robust speech recognition in severely adverse environments, where transcription of the adaptation data is not available.

To minimize the acoustic 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). Recently, missing-feature methods have shown promising results [1]-[5].

Acoustic model adaptation is generally considered to be one of the most popular approaches to effectively improve speech recognition in adverse environments. However, in real operating conditions, data for model adaptation may not be easily obtained, which represents the target condition for speech recognition. Effective transcription is also not available in general, which is usually required for conventional adaptation methods such as MAP [2] and MLLR [3]. In particular, such a situation can be easily encountered for the languages where research on speech technology is not extensively explored (i.e., low resource language).

In order to utilize untranscribed data for model adaptation, an unsupervised training technique is employed [6]-[8]. In general, a baseline ASR system is used for transcribing the adaptation data, and the resulting transcription is used for model adaptation. An iterative procedure is also employed, where the adapted model is used again for generating a more accurate transcription. For such an approach, the baseline ASR system is expected to be sufficiently reliable to generate a reasonably accurate transcription for unseen data, and some part of the adaptation data needs to be manually transcribed for an initial model. However, if mismatch between the baseline system and target condition representing the adaptation data is significant, the transcription performance drastically degrades, so the resulting model adaptation cannot be accomplished successfully.

In this study, we propose a simple model adaptation technique to increase speech recognition performance using untranscribed adaptation data. In the proposed method, a clean Gaussian Mixture Model (GMM) is obtained offline, and adapted over the adaptation data to generate an environment-dependent (i.e., noise-corrupted) GMM. Each Gaussian component of the clean Hidden Markov Models (HMM) is transformed using the mixture-selective transformation of the most similar Gaussian component of the GMM, resulting in a noise-corrupted HMM. Therefore, the proposed technique does not require transcription for the adaptation data. In the experiments, the proposed adaptation method is evaluated on the Linguistic Data Consortium (LDC) [9] RATS corpus, which represents severely adverse communication environments. To prove the effectiveness of the proposed method over an independent corpus, the LRE-07 and LRE-09 database [9] are also evaluated.

2. Communication Channel Environment Speech Corpus

This section describes the speech corpus used in this study. The speech corpus has been released by LDC as part of the Robust Automatic Transcription of Speech (RATS) program which is sponsored by the Defense Advanced Research Projects Agency (DARPA). The goal of the project is to create technology capable of accurately determining speech activity regions (SAD),

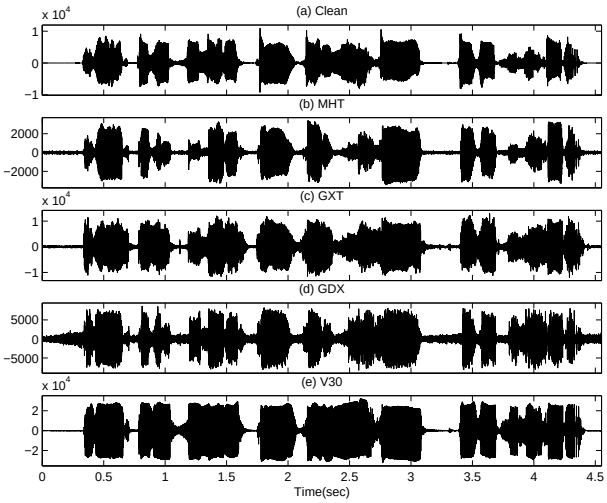

Figure 1: *Example of speech sample from different channels: (a) original clean, (b) MHT, (c) GXT, (d) GDX, and (e) V30.*

detecting key words (KWS), and identifying language (LID) and speaker (SID) in highly degraded communication channels. The speech corpora in the project are generated by transmitting an original clean speech database over different combinations of transmitter and receivers using LDC's multi radio-link channel collection system. The LDC Callfriend Farsi [9] corpus was used as source data for the speech corpus used in our study. The original Callfriend Farsi consists of western Farsi language conversations which were recorded over the telephone line.

In this study, four channels (MHT, GXT, GDX, and V30)[1] of data are selected, which reasonably represent actual field communication environments. They indicate different combination of transmitters and receivers, degrading original speech signals by different characteristics of signal modulation methods and carrier frequencies. The effects of the communication channels formulate distortions in speech signals, which are considered as convolutional interfering components. Fig. 1 presents an example speech utterance from the four channels (a) original clean, (b) MHT, (c) GXT, (d) GDX, and (e) V30 used in this study. As seen, there is only some background noise, however, the transmitted speech signals for the different channels contain considerably changes in waveforms structure. We believe that the different channel conditions bring highly different distortion effects to the signals, which result in significant mismatch between the original and transmitted speech signals. The objective speech quality measures including Signal-to-Noise-Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) are evaluated over the LDC RATS Farsi corpora, and details will be discussed in Sec. 4.

3. Gaussian Map Based Model Adaptation using Untranscribed Data

The proposed model adaptation method employs GMM adaptation using untranscribed adaptation data. A noise-corrupted speech GMM is obtained via a conventional adaptation technique, and used for a Gaussian mapping procedure to generate adapted HMMs. As an initial stage, a K-component GMM representing the clean speech signal $\mathbf{x}$ in the cepstral domain is estimated offline from the clean training data, which is given by,

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

In this study, $\mathbf{x}$ consists of 39 elements including the static feature vector (c0-c12) and dynamic feature vectors (i.e., first and second time derivatives).

Next, a noise-corrupted GMM is obtained by utilizing a conventional model adaptation method. In our experiment, MAP adaptation is employed, which generally provides improved performance compared to MLLR when the adaptation data is sufficiently available. As expected, transcription of the adaptation data is not required, since the adaptation is applied to the GMM. In this study, only mean and variance parameters are updated. The obtained noise-corrupted GMM can be represented by,

$$p(\mathbf{y}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}). \quad (2)$$

Now each kth Gaussian component $p(\mathbf{x}|k)$ of the clean GMM has a one-to-one mapping relationship with the corresponding component $p(\mathbf{y}|k)$ of the noise-corrupted GMM as follows,

$$\{\boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}\} \leftrightarrow \{\boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\}. \quad (3)$$

In the proposed model adaptation method, the most similar Gaussian component of the clean GMM is determined for each Gaussian component of the clean HMMs, and then the transform of the corresponding component in the noise-corrupted GMM is applied again to the Gaussian component of the HMM to generate noise-corrupted HMMs. The sth state's ith Gaussian component of the output probability function $q(\mathbf{x}|s,i)$ consisting of the clean HMMs can be represented by $\{\boldsymbol{\mu}_{\mathbf{x},s,i}, \boldsymbol{\Sigma}_{\mathbf{x},s,i}\}$. The proposed method employs a KL distance to measure the statistical similarity between Gaussian components,

$$k_{s,i}^{\min} = \arg\min_{k} \{\text{KL\_dist}(p(\mathbf{x}|k), q(\mathbf{x}|s,i))\}. \quad (4)$$

In the proposed method, the mean parameter $\boldsymbol{\mu}_{\mathbf{y},s,i}$ for the noise-corrupted HMMs is obtained by compensating a constant bias transform of the most similar Gaussian component of the clean GMM, and the variance $\boldsymbol{\Sigma}_{\mathbf{y},s,i}$ is generated by replacing one of corresponding components in the noise-corrupted GMM as follows,

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y},s,i} &= \boldsymbol{\mu}_{\mathbf{x},s,i} + (\boldsymbol{\mu}_{\mathbf{y},k_{s,i}^{\min}} - \boldsymbol{\mu}_{\mathbf{x},k_{s,i}^{\min}}) \\ \boldsymbol{\Sigma}_{\mathbf{y},s,i} &= \boldsymbol{\Sigma}_{\mathbf{y},k_{s,i}^{\min}}. \end{aligned} \quad (5)$$

The assumption of the constant bias transform of the mean parameter in the cepstral domain is motivated by other data-driven methods [10]. Fig. 2 illustrates the procedure of model adaptation in both the GMM model space and HMM model space. Here, the bold faced arrow and ellipse indicate the bias transform of the mean vector and variance of the determined Gaussian component, which are utilized to generate the Gaussian component of the adapted HMMs.

4. Experimental Results

4.1. Corpus Evaluation

To observe the degree of the channel corruption of the LDC RATS Farsi speech data used in this study, we evaluated objective speech quality measures including SNR and PESQ [11].

[1] They are originally symbolized as mht_aor, gxt_tr4, gdx_i75, and v30_v24, or channel A, C, D, and G respectively.

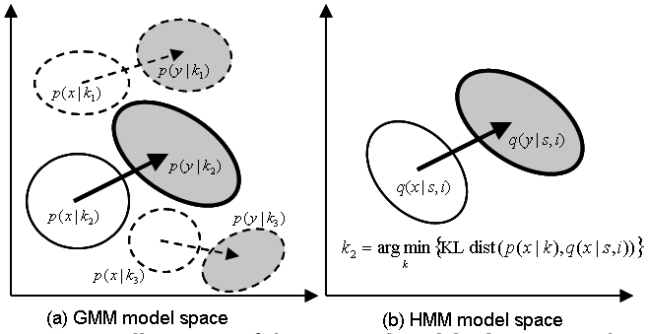

Figure 2: *Illustration of the proposed model adaptation technique: (a) One-to-one mapping of Gaussian components of clean and noise-corrupted GMMs, and (b) Adapted Gaussian component (s, i) of HMM utilizing the transform of the most similar GMM mixture component k_2.*

Table 1: STNR (dB) and PESQ (-0.5 to 4.5 MOS scale) over LDC RATS Farsi data.

	MHT	GXT	GDX	V30
STNR	23.42	26.21	21.84	59.23
PESQ	2.86	2.63	1.96	3.87

The Farsi data consists of 4565 segments from 17 audio recordings providing about 7-hour duration in total per each channel. The SNR was obtained using the NIST Speech Quality Assurance tool (i.e., the STNR estimator) [12]. Table 1 shows the averaged STNR and PESQ values measured over all segments per each channel. From the results, the channels GDX and V30 show the lowest and highest values in both STNR and PESQ measures respectively. For comparison, landline telephone NTIMIT and cellphone CTIMIT corpus show 35.79 dB/2.19 and 24.14 dB/1.74 in averaged STNR/PESQ respectively. These evaluation results suggest that the channel effects for the speech corpus used in this study bring significant corruption in signals, providing severely adverse environments for speech recognition.

4.2. Baseline ASR System Performance

Among the 4565 segments per each channel of Farsi data, 3651 and 914 segments were used for training/adaptation and testing respectively in these experiments. We employed SPHINX3 [13] as the Hidden Markov Model (HMM) based speech recognizer, which was built using training data from the clean channel. Each HMM represents a tri-phone which consists of 3 states with an 8-component GMM per state, which is tied with 2093 states. The task has 8765 Farsi words as the vocabulary, and a trigram language model was obtained by training on the transcription of the test set. A conventional MFCC feature front-end is employed in the experiment, which was suggested by the European Telecommunication Standards Institute (ETSI) [15]. An analysis window of 25 msec in duration is used with a 10 msec skip rate for 16-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 derivatives are also included, resulting in a feature vector of 39-dimension. The baseline ASR system shows 47.63% in Word Error Rate (WER) for the clean (channel) condition data.

Performance of the baseline system was evaluated over each of the channels using several existing robustness front-end algorithms. Spectral Subtraction (SS) [1] and Cepstral Mean Normalization (CMN) were selected as conventional algorithms. They represent some of the most commonly used techniques for additive noise suppression and removal of channel distortion respectively. The Advanced Front-End (AFE) algorithm developed by ETSI was also evaluated, which contains an iterative Wiener filter and blind equalization [16].

Table 2: Recognition performance in WER (%) employing conventional front-end algorithms with clean training and matched training conditions.

Clean Training	Clean Testing Condition: 47.63				
	MHT	GXT	GDX	V30	Avg.
CMN	87.31	90.66	95.56	71.98	86.63
SS + CMN	78.53	86.28	93.26	62.36	80.11
ETSI AFE	86.49	91.55	93.04	72.59	85.92
Matched Training SS + CMN	55.56	61.62	59.91	50.60	56.92

Table 2 shows speech recognition performance in WER using the conventional robustness front-end algorithms. It can be seen that combination of SS and CMN significantly improves recognition performance over all different channels. It is worth noting that the ETSI AFE was not effective for all channel conditions compared to the combination of SS and CMN, which is reported by many research groups to be highly effective at improving speech recognition accuracy in various noisy conditions. It indicates that the communication channel conditions included in the speech corpus used in these experiments are extremely challenging for speech recognition. The last row of Table 2 shows the recognition performance with a combination of SS and CMN in a matched training condition which was obtained by using the same channel degraded training data. It would provide an upper bound on performance for evaluating the model adaptation approaches in the next sections.

4.3. Performance Evaluation of the Proposed Adaptation Method using Untranscribed Data

Table 3 shows speech recognition performance employing various model adaptation methods using untranscribed data. The training portion (i.e., 3651 segments) per each channel is used as adaptation data, and their ground-truth transcription are assumed to be unavailable in this experiment. Here, both adaptation and testing data were processed using Spectral Subtraction and CMN, so the second row (SS+CMN) of Table 2 is a baseline performance for comparison. The transcription of adaptation data for MAP and MLLR was obtained in an unsupervised mode using the baseline ASR engine which is trained on clean channel data.

The results show that MAP adaptation was not effective compared to the baseline performance with 78.01% vs. 80.11% in averaged WER. Acoustic mismatch between HMMs of the baseline ASR and adaptation data should generate inaccurate transcriptions for model adaptation. It is also expected that the language model trained only on the original text of the test data decreases transcription performance further. Such a situation makes MAP performance ineffective, instead, MLLR shows relatively better performance with 72.33% in averaged WER, which transforms the model parameters using a matrix that represents an entire statistical change in the parameter space.

In the proposed Gaussian Map based Adaptation (GMA) method, the clean GMM with 512 Gaussian components is obtained by training over the training data of the clean channel, and the noise-corrupted (i.e., channel dependent) GMM is obtained by MAP adaptation of the clean GMM over the adapta-

Table 3: Recognition performance in WER (%) employing MAP, MLLR, and proposed GM-based adaptation algorithms.

	MHT	GXT	GDX	V30	Avg.
MAP	78.15	84.13	90.25	59.49	78.01
MLLR	65.62	76.97	86.39	60.34	72.33
GMA	65.89	78.32	81.75	52.39	69.59
GMA + MLLR	64.14	75.45	82.54	52.73	68.72

Table 4: Recognition performance in WER (%) employing MLLR and proposed GM-based adaptation algorithms using LRE-07 and LRE-09 database.

LRE-07	MHT	GXT	GDX	V30	Avg.
MLLR	68.61	80.23	89.15	67.59	76.40
GMA	67.82	79.31	83.51	55.63	71.57
LRE-09	MHT	GXT	GDX	V30	Avg.
MLLR	68.12	77.62	89.29	61.51	74.14
GMA	67.38	79.94	86.89	54.08	72.07

tion data per each channel. The proposed GMA method shows comparable performance to MLLR or outperforms MLLR, resulting in 69.59% averaged WER. In particular, GMA shows significantly improved recognition performance for channels GDX and V30. By combining MLLR as an initial model obtained by the GMA method, we obtained a slight improved performance for channels MHT and GXT. These results show that the proposed GMA adaptation method is significantly effective in severely noisy conditions, where transcribed data is unavailable and a baseline ASR fails to accurately transcribe the adaptation data due to acoustic condition mismatch.

4.4. Performance Evaluation using LRE-07 and LRE-09 Adaptation Data

In this section, we prove the effectiveness of the proposed adaptation method using independent adaptation data sets. Here, the adaptation data were generated by transmitting the Farsi language parts of LRE-07 and LRE-09 through the same radio-link channel collection system. A 2.6-hour and 4-hour set of speech data per each channel were used for LRE-07 and LRE-09 respectively. Table 4 shows speech recognition results using LRE-07 and LRE-09 data for model adaptation. Here it is also assumed that transcriptions of the adaptation data are not available. In a manner similar to the experiment of Table 3, the (clean condition trained) baseline ASR was used to generate the initial transcription for MLLR adaptation. Here, it can be expected that the scope of the vocabulary and language model as well as acoustic model mismatch would result in highly inaccurate transcriptions. The performance of MLLR in this experiment is lower compared to the results of MLLR in Table 3, since the mismatch of vocabulary and language model is larger in this experiment. It can be seen that the proposed GMA adaptation method does outperform conventional MLLR using both LRE-07 (71.57% vs.76.40%) and LRE-09 (72.07% vs. 74.14%) corpora as adaptation data. These results confirm that the proposed GMA method can be effectively employed to improve the acoustic model for speech recognition utilizing independent adaptation data sets without their corresponding transcriptions.

5. Conclusions

In this study, an acoustic model adaptation scheme was proposed to improve speech recognition in severely adverse environments utilizing untranscribed data. In the proposed method, a clean GMM is estimated from clean training data, and a noise-corrupted GMM is obtained by employing MAP adaptation over the adaptation data. The Gaussian component of the clean HMMs is transformed using the transformation of the most similar Gaussian component of the GMM, resulting in improved noise-corrupted HMMs. The proposed adaptation method was evaluated using the LDC RATS Farsi data which represents severely adverse communication channel environments. The method showed comparable or improved performance compared to conventional MLLR adaptation. The proposed method was also effective at improving speech recognition using independent adaptation data sets LRE-07 and LRE-09. The experimental results demonstrated that the proposed Gaussian Mapping based model adaptation method can be effectively employed to improve speech recognition, where transcribed data is unavailable and baseline ASR fails to accurately transcribe the adaptation data due to the acoustic condition mismatch.

6. Acknowledgment

This material is based on work partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024 (any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch), and by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

7. References

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

[2] J.L. Gauvain, 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, 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, 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] W. Kim, J.H.L. Hansen, "A Novel Mask Estimation Method Employing Posterior-Based Representative Mean Estimate for Missing-Feature Speech Recognition," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol.19, no.5, pp.1434-1443, July 2011.

[6] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *DARPA Broadcast News Trans. & Und. Workshop*, pp.301-305, Feb. 1998.

[7] L. Lamel, J.-L. Gauvain, G. Adda, "Unsupervised Acoustic Model Training," *ICASSP-2002*, pp. 877-880, May 2002.

[8] S. Novotney, R. Schwartz, J. Ma, "Unsupervised Acoustic and Language Model Training with Small Amounts of Labelled Data," *ICASSP-2009*, pp. 4297-4300, April 2009.

[9] http://www.ldc.upenn.edu

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

[11] Y. Hu, P. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. on Speech and Audio Processing*, vol.16, no.1, pp.229-238, 2008.

[12] http://www.nist.gov/speech.

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

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

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

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