Advanced Feature Normalization and Rapid Model Adaptation for Robust In-Vehicle Speech Recognition

Seong-Jun Hahm, Hynek Bořil, Pongtep Angkititrakul, and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Department of Electrical Engineering, Eric Jonsson School of Engineering & Computer Science
The University of Texas at Dallas, Richardson TX-75080, USA
{seongjun.hahm, hynek, pongtep.angkititrakul, john.hansen}@utdallas.edu

Abstract In this study, we present advanced feature normalization and rapid model adaptation for robust in-vehicle speech recognition. For feature normalization, we use a combination of recently established quantile-based cepstral dynamics normalization (QCN) and low pass temporal filtering (RASTALP). Similar to cepstral mean normalization (CMN), QCN aims at alleviating the mismatch between ASR acoustic models and the decoded speech signal. QCN relaxes CMN assumptions concerning feature distributions, making the normalization more stable in varying adverse environments. RASTALP is a low-pass approximation of RASTA filtering which significantly reduces transient distortions introduced by the original band-pass filter. Using the normalized features, we adapt the speaker-independent acoustic model to specific speakers. The adaptation method is based on an aspect model (a "mixture-of-mixtures" model). To enable adaptation requiring only extremely small amounts of adaptation data (i.e., a few seconds), we train a small number of mixture models which can be interpreted as models of probabilistic "speaker clusters" for in-vehicle environments. In this work, we use fMLLR to represent individual speaker models. The speaker models are mixed using weights determined from adaptation data. Experimental results show that the normalization employing QCN-RASTALP is consistently superior to CMN. We also observe that in contrast to the conventional methods, the adaptation based on the aspect model improves word error rates for the in-vehicle noise environments.

Keywords Robust speech recognition, in-vehicle environment, QCN-RASTALP, aspect-model-based adaptation

1. INTRODUCTION

Background noise is considered one of the most challenging problems for in-vehicle speech recognition systems. For solving this mismatch problem caused by varying car noises, many different methods have been proposed: These methods can be divided into three types: recording-based, analysis-based, and model-based methods. Among these approaches, we focus on the model-based method. In particular, we introduce advanced front-end feature normalizations and model adaptation for in-vehicle environment using small amount of adaptation data.

For feature normalization, we investigate several traditional schemes: cepstral mean normalization (CMN), cepstral gain normalization (CGN) [1], and recently established quantile-based cepstral dynamics normalization (QCN) [2] and a low pass temporal filtering (RASTALP) [3]. While conventional CMN and CGN are known to alleviate the effects of background noise and speaker/channel variability, their efficiency strongly depends on assumptions of Gaussianity or symmetricity of the sample distributions. QCN is designed to align dynamic ranges of the sample occurrences using quantile intervals estimated from the sample histograms and as such has no requirements on the shape of the sample distribution contours. QCN is combined with a temporal filtering strategy RASTALP that is inspired by the popular RASTA filter [4]. RASTALP approximates the low-pass portion of the RASTA band-pass filter by a second order Butterworth filter. The low order RASTALP significantly reduces transients effects seen in traditional RASTA [5]. Both separate and combined, QCN and RASTALP were previously found to provide superior normalization performance in noisy [5], Lombard effect [2], and reverberated [6, 7] speech recognition tasks. In addition, they displayed a competitive performance in the recent NIST SRE 2012 speaker recognition evaluation [8].

For model adaptation, there are two major approaches for the adaptation: Bayesian-based maximum a posteriori (MAP) [9] adaptation and the transform-based maximum likelihood linear regression (MLLR); unconstrained MLLR and constrained MLLR (equivalently fMLLR) [10–13]. The MAP approach is sufficiently stable for limited adaptation data (even though there is very little or no improvement). Maximum likelihood linear regression (MLLR) which is used for speaker adaptation, can easily be applied for noise adaptation. Although MAP and MLLR are effective for speech recognition in noisy environments, they require large amounts of adaptation data (more than 10 sentences; this could be worse in unsupervised adaptation case) for achieving sufficient performance.

In this paper, we introduce model adaptation based on aspect models. For in-vehicle environment ASR systems, the adaptation must be fast and rely on limited adaptation data. To realize a rapid adaptation, efficient approximation of inherent speaker/environment-specific characteristics is needed using extremely small number of adaptation data. The proposed Bayesian adaptation method exploits an aspect model: a "mixture-of-mixture" model. The difference from the previous work is that each speaker model is represented using speaker specific fMLLR transformation matrices (i.e., fMLLR adaptation in the training phase) instead of training speaker model using speaker specific data. In the presented framework, small numbers of "aspect models" are trained first based on maximum likelihood estimation, which are mixtures of distributions of the original unadapted model. When the adaptation data are given, the aspect models (i.e., basis models) are combined so that the likelihood for the adaptation data is maximized. The mixture weights are determined based on the maximum likelihood maximization (MLE). We evaluate the effectiveness of the proposed feature normalization and model adaptation methods by employing a speech recognition experiment under the in-vehicle environment.

2. FEATURE NORMALIZATION

The following cepstral normalizations are considered in our study: cepstral mean normalization (CMN) [14], cepstral gain normalization (CGN) [1], and recently proposed quantile-based cepstral dynamics normalization (QCN) [2] combined with RASTALP tempo-

ral filtering [3].

2.1. CMN

Cepstral mean normalization is a widely used technique compensationg for the speech signal variability in the cepstral domain [14]. The main focus of CMN is on convolutional distortions, however, it is partially effective also in reducing the effects of talking style variability and additive environmental noise [2]. In our study, CMN is applied on the speaker level, i.e., all speaker utterances are used to estimate the mean cepstral vector utilized in CMN. While this approach allows for a more accurate estimation of the cepstral means, it may not be available in online processing scenarios where utterance by utterance decoding is required. The other normalizations considered in this paper are applied in an online, per-utterance fashion.

2.2. CGN

Cepstral gain normalization, CGN [1], is similar to cepstral mean and variance normalization (CMVN). In CGN, the variance normalization is replaced by dynamic range (estimated from minimum and maximum samples) normalization. Both [1] and later studies [5] found CGN providing superior performance to CMVN in noisy ASR tasks.

2.3. QCN-RASTALP

Quantile-based cepstral dynamics normalization, QCN [2], is inspired by both CMVN and CGN; in QCN, the dynamic range of cepstral sample occurrences is estimated from histogram quantiles. In the subsequent step, the histograms are centered to the quantile mean and their variance is normalized to a unit inter-quantile interval:

$$c_{n,i}^{QCNj} = \frac{c_{n,i} - \left(q_j^{Cn} + q_{100-j}^{Cn}\right)/2}{q_{100-j}^{Cn} - q_j^{Cn}}, \tag{1}$$

where q_j^{Cn} and q_{100-j}^{Cn} are jth and $(100-j)$th quantile estimates in the nth cepstral dimension. The quantiles are estimated on the utterance-level from the dimension-wise cepstral sample histograms. QCN was shown to provide good performance in both small and large vocabulary tasks on neutral and LE speech in noisy conditions [2, 5] and reverberation [6].

Band-pass filtering in RASTA [4] eliminates low frequency components (including DC) as well as components varying faster than typical for speech. RASTA is typically applied either on the filterbank outputs in the spectrum domain or in the cepstral domain and has been found to improve ASR robustness in noise and reverberation. The high order of the original IIR band-pass RASTA filter tends to introduce transient distortions at the signal instances where the energy changes rapidly (e.g., beginning/end of speech islands). In [3], it was shown that these transients can be significantly reduced when replacing the band-pass by a low order low pass filter denoted RASTALP that approximates the characteristics of RASTA around the high cut-off region. The DC suppression can be realized separately by CMN or similar normalizations. The coefficients of the second-order infinite-impulse response (IIR) RASTALP filter

$$H(z) = \sum_{m=0}^{M} b_m \Big/ \sum_{n=0}^{N} a_n \tag{2}$$

with: $\mathbf{B} = [b_0, b_1, b_2] = [0.10408, 0.20816, 0.10408]$, $\mathbf{A} = [a_0, a_1, a_2] = [1, -0.90342, 0.31973]$, assuming a 10ms window step. In our study, the low-pass filter is combined with QCN, yielding a compensation called QCN-RASTALP.

3. MODEL ADAPTATION

In this section, we briefly review the related approaches first, then we explain the proposed aspect model.

3.1. Maximum a Posteriori Estimation

In most speech recognition systems using HMMs, the model parameters such as means and variances are estimated using maximum likelihood estimation (MLE). The formula for MAP adaptation of mean parameters as the follows:

$$\hat{\boldsymbol{\mu}} = \frac{N_a \boldsymbol{\mu}_a + \tau \boldsymbol{\mu}}{N_a + \tau}, \tag{3}$$

where $\hat{\boldsymbol{\mu}}$ is the updated mean, $\boldsymbol{\mu}_a$ is the mean of the adaptation data, $\boldsymbol{\mu}$ is the original mean, N_a is the number of available adaptation data and τ is a control variable determined empirically. The MAP method can be regarded as finding the optimal combination of existing data and adaptation data [9].

3.2. Feature Space Maximum Likelihood Linear Regression

fMLLR is widely used for feature space adaptation. We note that we only consider global feature transformation matrix in this work. The transformed feature vector $\hat{\boldsymbol{o}}(t)$ is given by

$$\begin{aligned} \hat{\boldsymbol{o}}(t) &= \boldsymbol{A}^{\mathcal{F}} \boldsymbol{o}(t) + \boldsymbol{b}^{\mathcal{F}} \\ &= \boldsymbol{W}^{\mathcal{F}} \boldsymbol{\xi}(t), \end{aligned} \tag{4}$$

where $\mathbf{A}^{\mathcal{F}}$ is a regression matrix, $\mathbf{b}^{\mathcal{F}}$ stands for a bias term, $\boldsymbol{W}^{\mathcal{F}}$ represents a transformation matrix, and $\boldsymbol{\xi}$ is an augmented feature vector. A more detailed explanation of the direct method can be found in [13].

3.3. Speaker Adaptation Using Aspect Model

3.3.1. Aspect Model Training

In the training phase, we train a reduced core set of aspect models so that the linear combination of these aspect models can approximate each of the speaker matrices.

First, let us consider estimating a basis model for a specific state. In our approach, each state has its own weighting. These basis models are used for estimating speaker specific aspect models. Definitions of symbols are as follows:

- $\lambda_{k,z}$: the first-level weighting of the k-th speaker and the z-th aspect basis
- $\xi_{k,z}$: the second-level weighting of the k-th speaker and the z-th aspect basis
- $\boldsymbol{o}(t)$: the t-th feature vector
- $\psi_k(\boldsymbol{o})$: a posterior probability for speaker k using the feature, $\boldsymbol{o}$, and corresponding transformation matrix, $\boldsymbol{W}^{(k)}$

where the function $\psi_k(\hat{\boldsymbol{o}})$ is represented as follows:

$$\psi_k(\boldsymbol{o}) = \sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{W}^{(k)}\boldsymbol{o}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m). \tag{5}$$

We keep the original GMMs as represented in Eq. (3). The original GMMs are expanded by the total number of speaker, K in the training data. We define the probability distribution function for speaker k and transformed feature vector $\hat{\boldsymbol{o}}$ as,

$$p(\boldsymbol{o}|\Xi_k, \lambda) = \sum_{z=1}^{Z} \sum_{k=1}^{K} \xi_{k,z} \lambda_{k,z} \psi_k(\boldsymbol{o}(t)), \tag{6}$$

where

$$\Xi_k = \{\xi_{1,1}, \cdots, \xi_{k,Z}\}. \quad (7)$$

The optimal $\lambda_{k,z}$ and $\xi_{k,z}$ can be found as [15, 16],

$$\lambda_{k,z} = \frac{\sum_t \alpha_{t,z}\beta_{t,k}}{\sum_k \sum_t \alpha_{t,z}\beta_{t,k}}, \quad \xi_{k,z} = \frac{\sum_{t:v_{t=k}} \alpha_{t,z}}{\sum_t \sum_{t:v_{t=k}} \alpha_{t,z}}, \quad (8)$$

where $\alpha_{t,z}$ and $\beta_{t,k}$ are the expectations of being in the z-th aspect model and k-th speaker models, respectively. This process is applied to all states sharing ξ_z.

3.3.2. Weighting Combination for Adaptation

For adaptation of the aspect models, a weighted combination is performed using all available adaptation data. For this, this expectation maximization (EM) algorithm is applied for estimating $\bar{\xi}_z$, which is the updated ξ_z. When the adaptation data $\boldsymbol{y}(1), \boldsymbol{y}(2), \cdots, \boldsymbol{y}(T)$ are given, $\bar{\xi}_z$ is calculated as,

$$\bar{\xi}_z^{(n+1)} = \frac{\sum_s \sum_t \bar{\xi}_z^{(n)} \sum_k \psi_{k,s}(\boldsymbol{y}(t))}{\sum_z \sum_s \sum_t \bar{\xi}_z^{(n)} \sum_k \psi_{k,s}(\boldsymbol{y}(t))}. \quad (9)$$

where n represents the number of iterations and s is the specific model state.

4. EXPERIMENTS

Speaker adaptation performance for the proposed method was evaluated using the *CU-Move* Corpus [17] in an unsupervised fashion.

4.1. Experimental conditions

The training data consists of 32556 utterances (prompt reading) from 98 speakers (46 males and 52 females) of a certain topic collected in vehicle (about 33.7 hours of speech data; we selected training speaker based on documentation about *CU-Move* corpus [18]). The test set consisted of 13760 utterances (9.98 hours, 6704 words) from 40 speakers (20 males and 20 females). Table 1 shows the experimental setup.

Table 1. Experimental setup

Sampling rate	16 kHz
Feature vector	MFCC + Energy + Δ + $\Delta\Delta$ (39 dims.) LDA + MLLT
Frame length	25 ms
Frame shift	10 ms
No. of categories	43 phonemes
HMM topology	Context-dependent 1577 (CMN) and 1612 (QCN) states $\approx$ 8 mixtures per state 3-state left-to-right HMM
Training method	ML Baum-Welch
Language model	3-gram
Vocabulary size	85,808
Perplexity	151.05
OOV rate	15.07 [1]

The acoustic model training, decoding, and the following acoustic model adaptation procedures were performed with the Kaldi speech recognition toolkit [11].

[1] We note that the OOV rate is high, and improved language models will be considered in the future.

4.2. Effect of Feature Normalization

The first round of evaluation experiments is focused on establishing a baseline ASR performance on the *CU-Move* task. Three approaches to front-end cepstral normalization (see Sec. 2) – CMN, CGN, and QCN-RASTALP (for space reasons denoted 'QCN') – are compared in Table 2. It can be seen that the order of performance is identical for setups with and without LDA and MLLT, with CGN reducing the WER of CMN and QCN outperforming both CMN and CGN. In both cases, QCN provides more than 1.3 % absolute WER reduction compared to CMN. The superior performance of QCN can be attributed to two factors – (i) more accurate alignment of the dynamic ranges of non-Gaussian cepstral distributions (low cepstral coefficients c_0–c_2 that reflect the energy and spectral slope of the speech signal tend to be multimodal and are very sensitive to the presence of environmental noise [2]) and (ii) the employment of RASTALP filtering that alleviates the impact of non-stationary noises [5].

Table 2. Word error rate comparison using CMN, CGN, QCN (%).

Model Training	Feature Normalization		
	CMN	CGN	QCN
MLE (MFCC)	40.57	40.50	**39.22**
MLE (LDA+MLLT)	38.61	37.39	**37.24**

4.3. Comparison with Existing Adaptation Methods

We performed experiments using conventional MAP, fMLLR, basis method [19], and the proposed aspect-model-based method. For MAP, we empirically set the control parameter τ to 20. For fMLLR, we used global diagonal full transformation matrix for the adaptation. For basis method, we set the number of coefficients d_n as same value in [19] (i.e., $\eta = 0.02$). In the proposed approach, we set the number of aspect models to 20. The experimental results are shown in Tables 3 and 4.

Experimental results show that all existing methods suffer from data sparsity problem when adaptation data is small. Even MAP and basis methods degrade performance which can be attributed to the unsupervised adaptation in this experiment. The proposed method did not degrade the performance for small adaptation data. In fact, the proposed method provided the best performance among all adaptations for small amounts of adaptation data (less than 15 seconds). Finally, it can be seen that all adaptation approaches consistently benefitted from applying QCN-RASTALP. The most dramatic WER reduction when switching from CMN to QCN-RASTALP can be observed in full-covariance fMLLR with absolute WER reduction by over 17 % for the smallest adaptation set. In addition, QCN-RASTALP increased the WER margin between the aspect model approach and MAP for the smallest adaptation set. These results demonstrate that the novel techniques presented in this paper do not only provide superior performance to their counterparts in the respective domains, but display also complementary benefits when combined together.

The small number of aspect models could adjust numerous parameters of reference speaker models. The advantage of the proposed method is that we can set different weightings for each state. However, the free parameters for adaptation are small because only ξ_z must be estimated in the adaptation phase (i.e., 20). We can also decide the set of aspect models using a regression class tree. This increases the number of free parameters, and therefore we can expect performance improvement according to the amount of adaptation data.

Table 3. Word error rate comparison using MAP, fMLLR, basis method, and the proposed basis method with CMN (%).

Adaptation Method	Baseline	Average amount of adaptation data (average length in frames and seconds)								
		500 (4.8)	1000 (8.1)	2000 (14.9)	5000 (36.4)	10000 (69.2)	20000 (135.4)	50000 (330.8)	100000 (646.7)	All (848.4)
MAP (20)	38.61	38.80	38.85	38.82	38.97	39.12	39.14	38.91	38.63	38.32
fMLLR (Full)		75.69	57.40	46.99	40.35	38.35	37.60	**36.80**	**36.51**	**36.36**
Basis		43.06	41.20	40.03	38.28	**37.67**	**37.37**	36.90	36.55	36.38
Aspect (20)		**38.23**	**38.23**	**38.23**	**38.22**	38.23	38.23	38.23	38.23	38.23

Table 4. Word error rate comparison using MAP, fMLLR, basis method, and the proposed basis method with QCN (%).

Adaptation Method	Baseline	Average amount of adaptation data (average length in frames and seconds)								
		500 (4.8)	1000 (8.1)	2000 (14.9)	5000 (36.4)	10000 (69.2)	20000 (135.4)	50000 (330.8)	100000 (646.7)	All (848.4)
MAP (20)	37.24	37.37	37.34	37.31	37.43	37.51	37.47	37.30	37.10	37.00
fMLLR (Full)		58.26	43.13	38.47	**36.77**	**36.31**	**35.98**	**35.71**	**35.49**	**35.40**
Basis		38.04	37.75	37.72	37.26	36.71	35.99	35.73	**35.49**	35.43
Aspect (20)		**36.78**	**36.78**	**36.78**	36.78	36.78	36.78	36.78	36.78	36.78

5. CONCLUSIONS

In this paper, we investigated advanced *feature normalization* and rapid model adaptation for robust in-vehicle speech adaptation. For feature normalization, we exploit advanced feature normalization, QCN-RASTALP, instead of CMN and CGN. For model adaptation, we used training information and a trained aspect model. Finally, we demonstrated performance gains using only weighted optimization. In the experiments, we confirmed that the proposed normalization and adaptation methods result in improved performance, even when the amounts of adaptation data was small.

6. REFERENCES

[1] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in *Proc. of ICASSP'04*, May 2004, vol. 1, pp. 209–212.

[2] Hynek Bořil and John H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, August 2010.

[3] H. Bořil, "UT-Scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," .

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on SAP*, vol. 2, no. 4, pp. 578 –589, Oct. 1994.

[5] H. Bořil, "Front-end compensation methods for LVCSR under Lombard effect," .

[6] O. S. Sadjadi, H. Bořil, and J. H. L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort," in *IEEE ICASSP'12*, Kyoto, Japan, 2012, pp. 4701–4704.

[7] H. Bořil, O. S. Sadjadi, and J. H. L. Hansen, "A study on combined effects of reverberation and increased vocal effort on asr," in *LISTA'12 Workshop*, Edinburgh, UK, 2012, pp. 16–19.

[8] T. Hasan, O. Sadjadi, L. Gang, N. Shokouhi, H. Bořil, and J. H. L. Hansen, "CRSS systems for 2012 NIST Speaker Recognition Evaluation," in *IEEE ICASSP 2013*, Vancouver, Canada, May 2013, pp. 6783–6787.

[9] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains," *IEEE Trans. on Speech and Audio Processing (SAP)*, vol. 2, no. 2, pp. 291–298, 1994.

[10] C.J. Leggetter and P.C. Woodland, "Speaker adaptation of HMMs using linear regression," *Cambridge University, Cambridge, UK, Tech. Rep. CUED/F-INFENG/TR*, vol. 181, 1994.

[11] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[13] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

[14] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[15] S. Hahm, Y. Ohkawa, M. Ito, M. Suzuki, A Ito, and S. Makino, "Aspect-model-based reference speaker weighting," in *Proc. of ICASSP*, 2010, pp. 4302–4305.

[16] S. Hahm, Y. Ohkawa, M. Ito, M. Suzuki, A. Ito, and S. Makino, "Improved reference speaker weighting using aspect model," *IEICE Trans. on Information and Systems*, vol. 93, no. 7, pp. 1927–1935, 2010.

[17] J.H.L. Hansen, P. Angkititrakul, P. Plucienkowski, S. Gallant, H. Yapanel, L. Pellom, W. Ward, , and A. Cole., "CU-Move: analysis and corpus development for interactive in-vehicle speech systems," in *Proc. of INTERSPEECH*, 2001, pp. 2023–2026.

[18] J.H.L. Hansen, "Getting Started with the CU-Move Corpus," Tech. Rep., 44 pgs, UTDallas-CRSS, Nov. 17, 2002.

[19] D. Povey and K. Yao, "A basis representation of constrained MLLR transforms for robust adaptation," *Computer Speech & Language*, vol. 26, no. 1, pp. 35–51, 2012.