SPEAKER-INDEPENDENT NAME RECOGNITION USING IMPROVED COMPENSATION AND ACOUSTIC MODELING METHODS FOR MOBILE APPLICATIONS

Kaisheng Yao, Lorin Netsch, and Vishu Viswanathan

Speech Technologies Laboratory, Texas Instruments
12500 TI Blvd, MS 8649, Dallas, TX 75243
{kyao, netsch, v-viswanathan}@ti.com

ABSTRACT

Name recognition is an important application of automatic speech recognition in embedded devices. Since embedded devices are used in diverse environments, noise robustness is very important. Moreover, unlike normal computer-based speech recognition applications, embedded speech recognition must deal with problems arising from limited resources. Facing these challenges, we have developed environment compensation and acoustic modeling techniques that improve robustness and accuracy of a speaker-independent name recognition system in hands-free conditions. These techniques are efficient to implement and are effective for performance improvement. On a name recognition task, we observed more than 53% word error rate reduction, compared to a baseline system. These improvements were obtained with minimal increase of resources.

1. INTRODUCTION

With the widespread use of mobile devices, automatic speech recognition (ASR) in embedded systems has become one of the major research and development areas for easy-to-use human machine interfaces. Since mobile devices are portable, ASR systems in such devices have to be robust and accurate in adverse acoustic environments. Likewise, due to limited computation and memory resources, these systems must be designed specifically to accommodate these limitations.

This paper describes compensation and acoustic modeling techniques in an ASR system for mobile applications. Compensation aims at reducing mismatch between trained acoustic models and real world testing speech signals. The mismatch is usually caused by background noise, for example, wind and car engine noise, and convolutive channel distortion such as hand-held versus hands-free microphones. Since the mismatch is subject to frequent change in mobile applications, it is critical to adaptively compensate the effects of the mismatch. Pursuing practical yet powerful compensation methods, our lab in the past has proposed a JAC method [1]. The method adapts mean vectors of acoustic models via a parametric mismatch function. To adaptively compensate frequently changing environmental distortion, a dynamic updating scheme is used in the method. However, performance of the method may be limited, especially in high noise levels. This paper presents an improved method, denoted as IJAC (Improved method of Joint compensation of Additive and Convolutive distortions) that modifies dynamic updating formulae to better deal with high noise levels. The modification is shown to reduce word error rate (WER) by 29%, compared to the JAC method.

Another key component in ASR systems is acoustic models. The widely used hidden Markov model (HMM) for acoustic modeling uses Gaussian mixtures to approximate the probability density functions (PDFs) of speech observations. To achieve high accuracy, HMMs are usually trained to be context dependent and have a large number of Gaussian PDFs. However, an embedded ASR system requires minimizing the number of Gaussian PDFs to conserve memory resources and computation. This paper describes a generalized tied mixture (GTM) scheme that effectively improves performance without increasing the number of Gaussian PDFs. Used together with the IJAC method, we observed more than 53% WER reduction compared to a baseline system.

2. THE COMPENSATION METHOD

Our compensation method belongs to model-space methods [1–3] that transform acoustic model parameters through a parametric function. Compared to feature-space methods such as speech enhancement, model-space methods usually yield better performance. Since model-space methods require fewer adaptation utterances for reliable transformation, they are more widely used for mobile applications, compared to some regression methods [4]. In the following, we briefly introduce the JAC method [1] and describe some improvements on it.

2.1. Environment compensation

We assume that noisy cepstral observation $Y(k)$ at time k is conditionally dependent on clean speech $X(k)$, background noise $N(k)$, and channel distortion $H(k)$. The clean speech cepstral observation $X(k)$ is assumed to be generated from a Gaussian PDF $\mathcal{N}(\cdot; \mu_{qp}, \Sigma_{qp})$ with mean vector μ_{qp} and diagonal covariance matrix Σ_{qp} at mixture p of state q. The PDF p has weight w_{qp} at state q of an HMM Λ_X.

Denote the log-spectral domain as superscript l. It is reasonable to assume that, within one utterance, environmental distortions, $N^l(k)$ and $H^l(k)$, are stationary. Denote their mean vectors in an utterance as $\Lambda_N = (N^l, H^l)$. We further assume a parametric mismatch function $g(\cdot)$ of environmental distortion of the clean speech mean vector as [1, 2]

$$\hat{\mu}_{qp} = Cg(\mu^l_{qp}, H^l, N^l), \tag{1}$$

where the parametric function $g(\cdot)$ is $g(X^l, H^l, N^l) = \log(\exp(X^l +$

$H^l)+\exp(N^l))$. C denotes the Cosine transformation. The likelihood of the noisy observation is therefore $p(Y(k)|qp,\Lambda_N,\Lambda_X) = \mathcal{N}(Y(k);\hat{\mu}_{qp},\hat{\Sigma}_{qp})$ for state q and mixture p [1]. In the following, we use $\mathcal{G}$ to denote $g(\mu^l_{qp},H^l,N^l)$.

2.2. Dynamic environment estimation process and its improvements

The objective of the improved method is to estimate the distortion Λ_N and to compensate its effect on clean speech models. Since Λ_N is subject to change between utterances, a dynamic estimation process is necessary. Denote the estimate in utterance u as $\Lambda_N^{(u)} = (N^{l(u)},H^{l(u)})$. The cost function for EM estimation of $\Lambda_N^{(u)}$ is $\mathcal{Q}(\Lambda_N^{(u)}|\bar{\Lambda}_N^{(u)}) = \sum_{c,k,q,p}\gamma_{qp}^{(c)}(k)\log p(Y^{(c)}(k)|qp,\Lambda_N^{(c)},\Lambda_X)$, which is proportional to $-\frac{1}{2}\sum_{c,k}\sum_{qp}\gamma_{qp}^{(c)}(k)\frac{1}{\hat{\Sigma}_{qp}^{(c)}}(Y^{(c)}(k)-\hat{\mu}_{qp}^{(c)})^2$. $\gamma_{qp}^{(c)}(k)$ is the posterior probability of being at state q and mixture p at frame k of utterance c. The compensation method iterates between obtaining the posterior probabilities given $\bar{\Lambda}_N^c$ estimated from the previous utterances and updating $\Lambda_N^{(u)}$ by maximizing the above cost function [2] for the current utterance u. In the following, we use $\mathcal{Q}$ to denote $\mathcal{Q}(\Lambda_N^{(u)}|\bar{\Lambda}_N^{(u)})$.

Background noise N^l is estimated by averaging non-speech frames of the current utterance. On the contrary, channel distortion usually varies less between utterances. Based on this observation, Newton's method is implemented for segmental updating of channel distortion in the JAC method [1], which involves the first- and second-order differentials of the cost function. Denote the first- and second-order differentials as $\Delta_{H^l}\mathcal{Q}$ and $\Delta^2_{H^l}\mathcal{Q}$. Direct calculation of the differentials involves transformations of variance $\hat{\Sigma}_{qp}^{(u)}$ between cepstral and log-spectral domain, which is computationally costly for mobile devices. To minimize implementation costs, the JAC method [1] adopted simplified differentials as

$$\Delta_{H^{l(u)}}\mathcal{Q} = -\sum_{c,k,q,p}\gamma_{qp}^{(c)}(k)[\mathcal{G}-C^{-1}Y^{(c)}(k)], \quad (2)$$

$$\Delta^2_{H^{l(u)}}\mathcal{Q} = -\sum_{c,k,q,p}\gamma_{qp}^{(c)}(k)\Delta_{H^l}\mathcal{G}, \quad (3)$$

where $\Delta_{H^l}\mathcal{G} = \frac{\exp(H^{l(u)}+\mu^l_{qp})}{\exp(H^{l(u)}+\mu^l_{qp})+\exp(N^{l(u)})}$ is obtained by referring to the parametric function $g(\cdot)$.

In this paper, we propose using the following simplified differentials of the cost function to achieve low computational costs and improved channel estimates.

$$\Delta_{H^{l(u)}}\mathcal{Q} = -\sum_{c,k,q,p}\gamma_{qp}^{(c)}(k)\Delta_{H^l}\mathcal{G}\left[\mathcal{G}-C^{-1}Y^{(c)}(k)\right] \quad (4)$$

$$\Delta^2_{H^{(u)}}\mathcal{Q} = -\sum_{c,k,q,p}\gamma_{qp}^{(c)}(k)[(\Delta_{H^l}\mathcal{G})^2 + (\mathcal{G}-C^{-1}Y^{(c)}(k))\Delta^2_{H^l}\mathcal{G}], \quad (5)$$

where $\Delta^2_{H^l}\mathcal{G}$ is the second-order differential of the parametric function $g(\cdot)$.

We may relate the proposed differential formulae with those in the JAC method [1] as

- removal of $\Delta_{H^l}\mathcal{G}$ from Eqs. (4) and (5),
- and assumption of $\exp(N^{l(u)}) \ll \exp(H^{l(u)}+\mu^l_{qp})$, i.e., additive noise power is much smaller than channel distorted speech power.

Because, in fact, the above assumption may not be valid, compensation by Eqs. (4) and (5) may perform better than JAC [1] particularly in high noise levels. We will confirm the statement through experiments in section 4.

Since mobile devices have constrained resource requirements, it is critical to verify that any improvement does not increase resource requirements significantly. Compared to the JAC method [1], the new method introduces approximately 5 more multiplications per mixture in Eqs. (4) and (5). Relative increase of computational costs is in fact very low, since 1) fixed-point DSP processors usually incorporate a hardware multiplier so that addition and multiplication can be completed in one CPU cycle, and 2) the differentials can be carried out only with the state-aligned models such that the actual number of involved HMMs is equal to just the length of the current utterance.

3. THE ACOUSTIC MODELING TECHNIQUE

As discussed in section 1, it is important to improve ASR performance without increasing the number of Gaussian PDFs in acoustic models. It is known that, given the same number of Gaussian PDFs, context-dependent acoustic models usually perform better than monophone models. Our baseline system uses intra-word triphones with a selected number of Gaussian PDFs for acoustic modeling. We developed the baseline HMM set such that, given the number, the single mixture per state system performed better than other systems with fewer states but larger number of PDFs per state. The observation suggested that we should design a training method that does not lose context-dependency details of the single mixture system.

The GTM scheme proposed in this paper employs a two-stage process to train HMMs. The first stage does the usual state-tying to train triphone models. State tying is achieved by the normal decision-tree-based state clustering. Triphone states are clustered according to their answers to questions in a phonetic binary tree with yes/no phonetic questions. A threshold is set to allow a certain depth of growing the decision tree to leaves that specify relevant contexts. We chose a threshold such that the single-mixture model achieved the highest performance.

In the second stage, Gaussian PDFs trained after state-tying are pooled together and eventually shared among different states and HMMs. The second stage in essence bootstraps the first stage. We use a statistic measure, the Bhattacharyya distance, to provide distances among PDFs; i.e. the distance between two Gaussian PDFs $\{\mathcal{N}_i(\cdot;\mu_i,\Sigma_i); i=1,2\}$ is

$$D(\mathcal{N}_1,\mathcal{N}_2) = \frac{1}{8}(\mu_1-\mu_2)^2(\frac{\Sigma_1+\Sigma_2}{2})^{-1}+\frac{1}{2}\ln\frac{(\Sigma_1+\Sigma_2)/2}{\|\Sigma_1\|^{1/2}\|\Sigma_2\|^{1/2}}$$

Given a PDF, sharing of PDFs can be done among Gaussian PDFs with the shortest distances to the given PDF. The idea of PDF sharing is illustrated in Fig. 1, where sharing of PDFs is among similar phones such as `ax` and `er`. Ability to discriminate phones is attained by 1) using different mixture weights, and 2) sharing different mixture PDFs with other states.

After the above process, each state of the HMMs has M Gaussian PDFs but the total number of Gaussian PDFs is kept

[1] Usually the new variance $\hat{\Sigma}_{qp}$ differs from the original Σ_{qp}.

[2] A forgetting factor $\rho \in (0,1.0]$ may be introduced to force parameter updating with more emphasis on recent utterances.

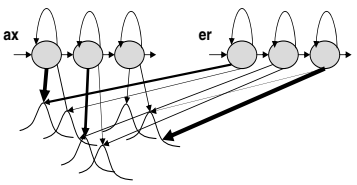

Fig. 1. Sharing Gaussian PDFs between `ax` and `er`.

unchanged. HMMs are then re-trained with flat initialization of mixture weights. Then, training is done on these mixture weights and transition probabilities. Finally, all HMM parameters are re-trained with several iterations of the Baum-Welch EM algorithm [5].

The GTM process is different from some alternative mixture tying methods. Compared to a pure mixture tying system such as semi-continuous HMM [6], the GTM-HMM uses state tying to preserve the state identity. Compared to the sole state tying system [7], such models share Gaussian PDFs across states even though these states may belong to different models. Instead of using phonetic knowledge to tie mixture PDFs such as that in [8], GTM uses a statistic measure which may "break" the constraint set by the phonetic knowledge. For example, using the GTM process, Gaussian PDFs in a speech HMM may be shared together with those PDFs in a silence model.

4. EXPERIMENTAL RESULTS

4.1. Database and baseline performance

Our ASR system was tested on a hands-free speech recognition database. The database was recorded in vehicles, using an AKG M2 hands-free distant talking microphone, in three recording sessions: parked (car parked, engine off), city-driving (car driven on a stop and go basis), and highway (car driven on a highway). In each session, 20 speakers (10 male/female) read 120 pairs of English first and last names. The database was sampled at 8 kHz, with frame rate of 20 ms. From the input speech, 10-dimensional MFCC features, together with their first-order differentials, were derived. The Wall Street Journal (WSJ) database was used to train the acoustic models. The baseline used the JAC method for environment compensation and single-mixture per state triphones for acoustic modeling. The average WER over the three driving conditions was 3.2%.

Further performance improvements must handle the following mismatches. First, the microphone is distant talking and band-limited in the database, compared to the high-quality microphone used to collect the WSJ database. Second, there is a substantial amount of background noise in cars, with SNR decreasing to 0dB in the highway condition. Third, utterances in the WSJ database are more continuous than the name utterances in the database.

4.2. Channel distortion estimates

We show in this section that the IJAC method has more reliable channel estimates than those of the JAC method. Figures 2 and 3 plot the mean and standard deviation of the estimated H^l by

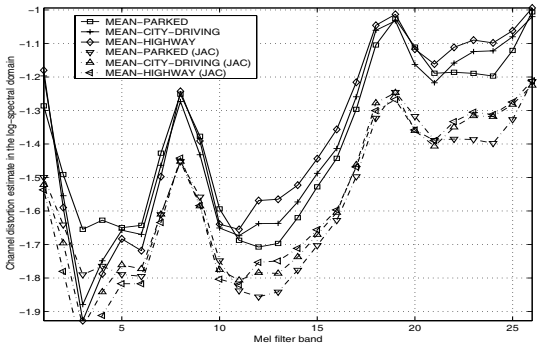

Fig. 2. Mean of channel distortion estimates averaged over all testing utterances.

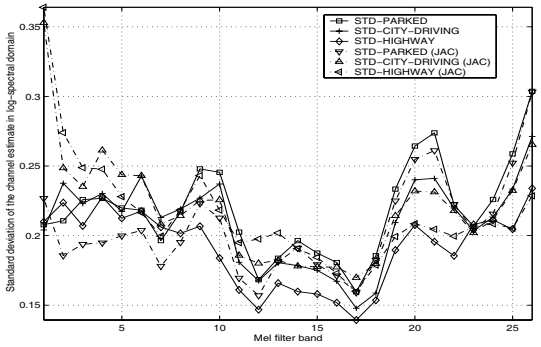

Fig. 3. Standard deviation of channel distortion estimates averaged over all testing utterances.

IJAC in the three driving conditions, averaged over each condition, together with those for the JAC method. From Fig. 2, we observe that, given a method, the estimates in different driving conditions are generally in agreement. However, Fig. 3 shows that the estimation variance of these methods are different. Whereas these two methods had similar estimation variances in higher frequency, they performed differently in lower frequency. Particularly, in highway and city-driving conditions, estimation variance of JAC was much larger than that in the parked condition. On the contrary, the IJAC method performed stably without much difference in estimation variances for the three driving conditions.

4.3. Recognition results

The IJAC method was compared to JAC [3] and MLLR [4]. Diagonal linear transformation matrices in MLLR were clustered using a binary phonetic tree. Forgetting factor ρ was set to 0.6 in both JAC and IJAC. Acoustic models were trained by the GTM process in section 4.4 with $M = 10$. The recognition results are summarized in Table 1. We observe that

- Without noise robustness techniques, the system performance degraded severely under the noisy environments found

[3] JAC performed better than PMC [2] in [1]. We therefore did not include results of PMC in this paper.

in the testing database. We found that most of the errors occurred in high noise levels.

- MLLR effectively improved noise robustness. Compared to the system "W/O compensation", WER decreased to 22.7%. The average of the three driving conditions indicates relative WER reduction was 50%.
- In mobile applications of ASR, the JAC method was more effective than MLLR. Using JAC, we further decreased WER to 2.1%, corresponding to 90% WER reduction averaged over the three driving conditions. Careful analysis of the experimental results showed that much of the performance improvement was achieved in the highway driving condition. We found that, due to high noise levels and frequent change of environments in the highway driving condition, it was very difficult for MLLR to reliably estimate a set of linear transformations for environment compensation. On the contrary, the estimation of background noise $N^{l(u)}$ and channel distortion $H^{l(u)}$ in JAC performed more stably than the estimation of linear transformations in MLLR. In the highway driving condition, we observed more than 92% relative WER reduction.
- Whereas the JAC algorithm substantially reduced WERs compared to "MLLR" and "W/O compensation", in all driving conditions, IJAC performed even better. The averaged WER decreased to 1.5%, corresponding to 29% relative WER reduction. We analyzed relative WER reductions in each of the driving conditions, and found that the relative WER reduction was more than 30% in the highway condition. The results clearly show that IJAC is very effective in compensating environment distortions.

Table 1. WER (in %) achieved by different methods on the hands-free task.

Methods	WER (in %)
W/O compensation	45.1
MLLR	22.7
JAC	2.1
IJAC	1.5

4.4. Acoustic modeling

This section presents results with the improved acoustic modeling technique described in section 3. We varied M, the number of sharable Gaussian PDFs per state, but kept the total number of Gaussian PDFs unchanged. Table 2 shows the results of the GTM process. We observe that, generally, larger M can decrease WER relative to that with $M = 1$. For example, using $M = 10$, we decreased WER from 2.2% by $M = 1$ to 1.5%, providing 31% relative WER reduction. We conducted careful analysis and found that such performance improvement was in particular consistent in the highway condition. We also compared the GTM process with other training schemes [8] and found that it performs better especially with larger M [9].

Referring to Tables 1 and 2, we observe significant performance improvement relative to our baseline system described in section 4.1. More than 53% WER reduction was achieved by combining the improved compensation and acoustic modeling techniques. Notice that the improvement was obtained without much increase of computation resources and foot-print.

Table 2. WER (in %) achieved by the GTM process with different M on the hands-free task.

M =	1	2	4	10
WER (in %)	2.2	2.0	1.7	1.5

5. CONCLUSIONS

We have presented in this paper improved methods of environment compensation and acoustic modeling. These methods aim at robust and accurate name recognition for mobile applications. Since embedded devices require limited resources, these methods are designed carefully to keep computation low and minimize foot-print. On a hands-free name recognition task, we obtained significant performance improvements using these improved methods, compared to a baseline system.

6. REFERENCES

[1] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 975–983, 2005.

[2] M.J.F.Gales and S.J.Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.

[3] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *ICASSP*, 1997, pp. 835–838.

[4] P. C. Woodland, M. J. F. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," in *ICASSP*, 1996, pp. 65–68.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[6] X.D. Huang, Y. Grilui, and M.A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[7] S. Young, *The HTK BOOK*, Cambridge University, 2.1 edition, 1997.

[8] Y. Liu and P. Fung, "State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition," *IEEE Trans on Speech and Audio Processing*, vol. 12, no. 4, pp. 351–364, 2004.

[9] K. Yao, "Generalized tied mixture hidden Markov models for automatic speech recognition," Tech. Rep. KY-2004-0004, Speech Technologies Laboratories, Texas Instruments, 2004.