

# Methods for Task Adaptation of Acoustic Models with Limited Transcribed In-Domain Data

Enrico Bocchieri, Michael Riley<sup>†</sup> and Murat Saraclar

AT&T Labs-Research, 180 Park Ave,  
Florham Park, NJ 07932, USA

{enrico,murat}@research.att.com riley@google.com

## Abstract

Application specific acoustic models provide the best recognition accuracy, but they are expensive, because they require the transcription of tens or hundreds of hours of in-domain speech for training. Therefore, this paper focuses on the acoustic model estimation given *limited* in-domain transcribed speech data, and large amounts of (typically available) transcribed out-of-domain data. First, we evaluate several combinations of known methods to optimize the adaptation/training of acoustic models on the limited in-domain speech data. Then, we propose to use Gaussian *sharing* to combine in-domain models with out-of-domain models, and a data generation process to simulate the presence of more speakers in the in-domain data. In a spoken language dialog application, we contrast our methods against an upper accuracy bound of 69.1% (model trained on many in-domain data) and a lower bound of 60.8% (no in-domain data). Using only 2 hours of in-domain speech for model estimation, we improve the accuracy by 5.1% (to 65.9%) over the lower bound; data generation and Gaussian sharing contribute 2.2% to this improvement. With 9 hours of in-domain speech, the improvement of accuracy is 6.5%, to 67.3%.

## 1. Introduction

In the field of acoustic modeling for speaker independent speech recognition, task independence is a long pursued, but still elusive goal. Many large, phonetically-rich, transcribed speech data bases for acoustic model training are currently available. Still, best recognition performance on a particular application requires acoustic models trained on tens or hundreds of hours of application specific (in-domain) speech data. Methods for porting a recognition system to a new task have recently been studied in [1].

Another consideration is that the cost of collecting and transcribing in-domain data limits the practical deployment of application-specific models. Therefore, lightly supervised [2], [3] and unsupervised [4], [5] training algorithms have been studied.

This work concentrates on a different scenario, with:

- (a) *Many* transcribed out-of-domain speech data, useful for training an acoustic model for the initial deployment of a new application, and
- (b) *Few* in-domain training data, as transcribed by a technician in few days work, after initial deployment.

We study how to best employ the in-domain data (b) to improve the performance of the acoustic model in (a). We evaluate standard model adaptation methods of the out-of-domain model on the in-domain data. We investigate how to fully re-train new models on the in-domain and out-of-domain data, to achieve the best compromise between sharpness and generalization of the model.

We introduce two novel ideas. The first idea is sharing output densities between in-domain and out-of-domain acoustic models. The second, speech data generation, transforms the in-domain

speech data to simulate production of the same sentences by different speakers. It artificially creates a larger in-domain training set, which leads to the estimation of more accurate models.

## 2. Methods

The recognizer frontend produces 39 dimension feature vectors, consisting of 12 mel frequency cepstrum coefficients (*MFCCs*), frame energy, and first and second differentials, at a frame rate of 100 per second. Mean subtraction is applied to the cepstrum features. Acoustic hidden Markov models (HMM) are context dependent (triphones) with tied states. State tying is performed by classification and regression trees [6]. HMM state output densities are Gaussian mixture models (GMM), estimated by maximum likelihood:

$$P(\mathbf{x}|s) = \sum_{i=1}^{N_s} p(\omega_i|s) \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{s,i}, \Sigma_{s,i}) \quad (1)$$

### 2.1. Adaptation

We investigate the performance of hidden Markov model adaptation to data from a new application by means of maximum likelihood linear regression (MLLR) [7], and maximum a posteriori criterion (MAP) [8]. In MLLR adaptation, linear transformations  $A_g$  are applied to the Gaussian means:

$$\boldsymbol{\mu}_{s,i,mllr} = A_g \boldsymbol{\mu}_{s,i} \quad (2)$$

where subscript  $g$  denotes the regression class for state  $s$ . MAP adaptation has been applied to the Gaussian means and covariances. For the means:

$$\boldsymbol{\mu}_{s,i,map} = \frac{\gamma m_{s,i} \boldsymbol{\mu}_{s,i} + n_{s,i} \bar{\mathbf{y}}_{s,i}}{\gamma m_{s,i} + n_{s,i}} \quad (3)$$

where  $\bar{\mathbf{y}}_{s,i}$  and  $n_{s,i}$  are the sample mean and count of the adaptation data,  $m_{s,i}$  is the Gaussian component count, and  $\gamma$  is related to the distribution prior. A similar expression holds for the adaptation of the second order statistics.

The language model is a word trigram language model, discussed further in Section 3.

### 2.2. Sharing Output Densities

Another method for making use of in-domain data is to build a new acoustic model only on the in-domain data and merging this in-domain model with the out-of-domain model. Here we use a technique that allows sharing of output densities among two acoustic models. The mechanism of sharing output densities among HMM states is inspired by *soft clustering* [10] for robust acoustic model estimation. This idea was extended to sharing densities across different acoustic models for pronunciation modeling [11]. Under this

<sup>†</sup>Now with: Google Inc., 1440 Broadway, New York, NY 10018, USA.

framework, the output density of a state in an acoustic model is enhanced by including Gaussian densities from the states of another acoustic model.

Given a base acoustic model  $A$  with a set of states  $\mathcal{S}_A$  and another acoustic model  $B$  with a separate set of states  $\mathcal{S}_B$ , the output density of each state  $s \in \mathcal{S}_A$  is modified as follows.

$$P(\mathbf{x}|s) = \lambda P_A(\mathbf{x}|s) + (1 - \lambda) \sum_{s' \in \mathcal{S}_B} p(s|s') P_B(\mathbf{x}|s') \quad (4)$$

where  $\lambda$  is an interpolation constant (typically 0.5) and  $p(s|s')$  is the conditional probability of state  $s$  of model  $A$  given state  $s'$  of model  $B$ . This conditional probability is estimated from a pair of state-level time alignments of the in-domain data obtained via forced alignment using the two models. The state-level time alignment assigns each acoustic frame to a state, and the pair of alignments provides an indirect alignment between the states of the two acoustic models. Let  $C(s, s')$  be the number of frames  $s$  aligned with  $s'$ , and  $C(s')$  be the total number of frames aligned with  $s'$ . Then the maximum likelihood estimate for the conditional probability is given by

$$p(s|s') = \frac{C(s, s')}{C(s')}.$$

In order to limit model size and for robustness we use a threshold on  $C(s, s')$  and another on  $p(s|s')$ . For pairs which are below these thresholds  $p(s|s')$  is set to zero.

Note that this method requires two acoustic models and time alignment of the in-domain data with these models. One of the models (Model  $A$ ) defines the structure and provides mixture components while the other (Model  $B$ ) only provides mixture components. Although the number of parameters is increased for the merged model, the robustness of the model is not affected since the component densities are not reestimated after merging. In fact, under certain conditions reestimating the model parameters can improve performance [11].

### 2.3. In-Domain Speech Data Generation

Many of the triphones observed in small training sets may be spoken by very few individuals, therefore, the corresponding triphonic models do not generalize well. To attenuate such a problem, we simulate the presence of more speakers through the application of speaker transformations of the in-domain speech. Vocal tract length is known to be a source of speaker variability that is important for automatic speech recognition. Changes of vocal tract length are modeled by a linear warping function of the speech frequencies:

$$f_{warp} = \alpha f \quad (5)$$

(to insure invertibility, we use a piece-wise linear warp [9]). We apply (5) in the *MFCC* computation for different values of  $\alpha$ , to simulate the pronunciation of the in-domain training sentences by individuals of different vocal tract length, respectively.

## 3. Experimental Setup and Results

We recorded the out-of-domain and in-domain data from two different spoken-language customer assistance applications, here denoted as "operator services" (*OpServ*), and "customer care" (*CustCare*), respectively [12]. The full training data sets of the acoustic models consist of 30,889 sentences (word count of 229,000) for *OpServ*, and 67,181 sentences (word count of 1,246,000) for *CustCare*. The dictionaries have sizes (number of words occurring at least once in the training data) of 3,336 (*OpServ*) and 12,563 (*CustCare*), and an intersection of 2,084 words.

# in-domain sentences	Word accuracy	
	MLLR	MAP
1,500	61.8%	62.4%
6,000	61.9%	64.9%

Table 1: Adaptation of out-of-domain HMM on in-domain data.

In practice, when porting a recognition system from an old (*OpServ*) to a new (*CustCare*) domain, one is confronted with mismatched lexicon, language and acoustic models. We want to isolate the acoustic model mismatch problem to study the performance of acoustic model training/adaptation algorithms. Therefore, all our experiments use the same lexicon and trigram language model built on 38,000 *CustCare* sentence transcriptions.

The test set consists of 667 sentences (10,875 words) from *CustCare*. The in-domain HMM, trained on *CustCare* data contains 4,116 states and 32,991 Gaussians. The out-of-domain HMM, trained on *OpServ*, contains 2,391 states and 19,160 Gaussians. Word accuracies for in-domain and out-of-domain acoustic models are 69.1% 60.8%, respectively. In our scenario, the accuracy loss of mismatched acoustic models is 8.3%.

In the next Sections we examine various methods to improve over the accuracy of *OpServ* HMM, given a relatively small number of transcribed sentences from *CustCare*. We consider two scenarios, with either 1,500 in-domain sentences (3.5 hours of audio,  $\approx$  2 hours of speech), or 6,000 in-domain sentences (14.4 hours of audio,  $\approx$  9 hours of speech), especially concentrating on the first case.

### 3.1. Out-Of-Domain Model Adaptation

We first performed adaptation of the *OpServ* acoustic model on the in-domain data (*CustCare*) by means of MLLR and MAP. The in-domain data is relatively abundant for MLLR adaptation, therefore we estimate a linear transform  $A_g$  (2) for every monophone. For MAP, we have set the constant  $\gamma$  in (3) by experiment. The error rates were not sensitive to the choice of  $\gamma$ . The results are summarized in Table 1. MAP adaptation outperforms MLLR, as expected considering the amount of in-domain adaptation data.

### 3.2. Retraining With Available In-Domain And Out-Of-Domain Data

We assume that the out-of-domain (*OpServ*) data can be combined with the few *CustCare* data to estimate a better model for the *CustCare* application. Intuitively, the combination of in-domain and out-of-domain data yields a compromise between improving the generalization capability of the HMM trained on the in-domain data only, and a reduction of the resolution of the HMM. The best balance depends on the amount of in-domain data and the number of parameters to estimate. We have considered three possibilities:

- (i) Use the in-domain data only, to build the context dependency trees of the HMM and for state output density estimation.
- (n) Build the context dependency trees with the in-domain data, and estimate the state output densities with the combined in-domain and out-of-domain data. In tree building, node splitting is based on a simple Gaussian model, with far fewer parameters than the GMMs. Therefore, our conjecture is that the in-domain data may be sufficient for tree building, but not for state output density estimation.
- (a) Generate both the context-dependency trees and the states using both in-domain and out-of domain-data.

# In-domain sentences, tree type and size	Number of states	Number of Gaussians	Word accuracy %
1,500 (i) small	940	7157	60.6%
1,500 (i) med	1,598	8,628	59.9%
1,500 (i) big	3,243	8,897	57.0%
1,500 (n) small	938	7530	62.0%
1,500 (n) med	1,612	12,617	<b>63.4%</b>
1,500 (n) big	3,278	24,048	62.8%
1,500 (a)	4,088	32,496	<b>63.7%</b>
6,000 (i)	1,643	13,169	<b>66.1%</b>
6,000 (n)	1,666	13,371	65.2%
6,000 (a)	3,634	30,772	65.6%

Table 2: Models trained with in-domain sentences.

# In-domain sentences, tree type and size	Word accuracy %
1,500 (i) small	62.7%
1,500 (i) med	62.5%
1,500 (i) big	60.0%
6,000 (i)	66.2%

Table 3: In-domain HMMs map-adapted on out-of-domain data

For training the context dependency trees on 1,500 in-domain sentences, we experimented with three tree sizes, denoted as "small", "medium" and "big", by allowing for a minimum count at tree leaves of 500, 250 and 100 respectively. The maximum number of Gaussian mixture components was set to 8 per state (16 for silence hmm), with a count of at least 50 samples for a Gaussian component.

The experiment results are shown in Table 2. With 1,500 *Cust-Care* sentences, model "1,500 (a)" in Table 2, gives the best result. However, model "1,500 (n) med", is much smaller with only slightly lower accuracy. These accuracies are better than those of MAP adaptation of the previous Section. With 6,000 in-domain sentences, the best strategy (66.1% accuracy with model "6,000 (i)" in Table 2) is simply to retrain the HMM with the in-domain data only.

### 3.3. Map Adaptation Of Retrained Models

It is evident from Table 2, case (i), that the HMM estimated on only 1,500 in-domain sentences do not generalize well. Therefore we attempted to improve the accuracy of the (i) models by smoothing the estimated state output densities with the available out-of-domain data, through MAP adaptation on the out-of-domain data (reverse adaptation). This step considerably improves the accuracy of the (i) models trained on 1,500 sentences, as shown in Table 3. However the best results are still obtained with models "1,500 (a)" and "1,500 (n) med" of Table 2.

When estimating the state output densities of acoustic model (n), the count of in-domain feature vectors is much smaller than for out-of-domain vectors, and the statistics may be biased in favor of the out-of-domain task. To attenuate such an effect we have adapted (MAP) the models (n) with in-domain data, thus increasing the relative weight of the in-domain data. Results are in Table 4. Overall, there is an improvement, but not over the best case with 1,500 sentences, in Table 2.

# In-domain sentences, tree type and size	Word accuracy %
1,500 (n) small	62.8%
1,500 (n) med	63.3%
1,500 (n) big	63.3%
6,000 (n)	65.7%

Table 4: Map adaptation of in-domain HMMs on in-domain data

### 3.4. Sharing output densities

We used the method described in Section 2.2 to merge in-domain and out-of-domain acoustic models. Recall that one of the models (Model A) defines the structure and provides mixture components while the other (Model B) only provides mixture components. We have experimented with various combinations of in-domain and out-of-domain acoustic models and the results are reported in Table 5. For the best results the gain over the previous best performing models are 1.5% (63.7% to 65.2%) for 1,500 utterances and 1.2% (66.1% to 67.3%) for 6,000 utterances.

Model A	Model B	Word Accuracy
out-of-domain	1,500 (i) small	<b>65.2%</b>
1,500 (i) small	out-of-domain	64.8%
out-of-domain	1,500 (i) med	64.4%
1,500 (i) med	out-of-domain	64.4%
out-of-domain	1,500 (i) big	64.1%
1,500 (i) big	out-of-domain	64.2%
out-of-domain	6,000 (i)	66.4%
6,000 (i)	out-of-domain	<b>67.3%</b>

Table 5: Sharing output densities between in-domain and out-of-domain models.

For small in-domain model sizes it is better to use the out-of-domain model as the base model. As the in-domain model gets larger this trend is reversed.

This technique results in a substantial increase in the number of parameters. As explained in Section 2.2 two thresholds are used to limit the size of the model. In our experiments both thresholds were set to 0.1, which results in having an average of 1.6 to 2.2 states from Model B sharing densities with each state in Model A.

It is important to note that for best results the models being merged should be trained on different data. Sharing of densities is not as effective if the densities are not significantly different. For example, merging a MAP adapted in-domain-model with an out-of-domain model yields a very small gain. In fact, the improvement from this technique is more when merging models with different context dependency structure.

### 3.5. Data Generation

For every transcribed in-domain sentence, we have artificially generated six more *MFCC* sequences by means of the frequency warp technique of Section 2.3, with six values of  $\alpha$ :

$$\alpha = 0.85, 0.9, 0.95, 1.05, 1.1, 1.15 \quad (6)$$

We have applied the techniques of the previous Sections, after augmenting the in-domain sentences with these artificial data. The HMM estimation parameters were chosen so that the model estimated with the additional artificial data have size nearly equal to those estimated on only the original data. Table 6 summarizes the experiments for 1,500 in-domain sentences. Both the adaptation and re-training results are significantly improved by the artificial

Improvements of accuracy (same model sizes) with data generation, using 1,500 in-domain sentences			
"1,500 MAP" from Table 1	62.4%	⇒	63.6%
"1,500 (i) med" from Table 2	59.9%	⇒	63.3%
"1,500 (n) med" from Table 2	63.4%	⇒	65.0%

Table 6: Data generation from 1,500 in-domain sentences.

data. The previous best accuracy of 63.7% is improved by 1.3% to 65.0%. The best accuracy with 1,500 in-domain sentences, was obtained by combining Gaussian sharing and data generation: merging the model that yields an accuracy of 63.6% in Table 6 with the out-of-domain model, gives an accuracy of 65.9%.

We can expect that more in-domain speech provides a better characterization of vocal tract length variations than just 1,500 sentences, thus making the addition of frequency warped data relatively less effective. In fact, in the case of 6,000 in-domain sentences, we obtained an accuracy improvement of only 0.2%, to 66.3% from 66.1% of Table 2.

## 4. Discussion and Conclusion

In general, when porting to a new application, one must decide whether to use an out of domain model or adapt or retrain on some in-domain data. For increasing amounts of in domain data  $t$ , suitable approaches are:

- (i) Use out-of-domain HMM, if  $0 < t < t_{mltr}$ .
- (ii) Apply MLLR adaptation to out-of-domain HMM, if  $t_{mltr} < t < t_{map}$ .
- (iii) Apply MAP adaptation if  $t_{map} < t < t_{ctx}$ .
- (iv) Retrain HMM context tree on in-domain data, and HMM states on in-domain and out-of-domain data, if  $t_{ctx} < t < t_{new}$ .
- (v) Retrain HMM (context tree and mixtures) on in-domain data, if  $t_{new} < t$ .

In our specific scenario with 6,000 in-domain sentences (v) gives the best results. Instead, with 1,500 in-domain sentences, (iv) and (v) provide similar accuracies. Therefore, in our scenario,  $t_{map}$  is less than 1,500 sentences, and  $t_{ctx}$  is in the neighborhood of 1,500 sentences. However, we believe that the different speech data quantities, that are best suited for the various methods (i) through (v), depend on the applications, like on the similarity between the triphonic contents of the in-domain and out-of-domain data. In fact, it appears that in the experiments of [1],  $t_{map}$  and  $t_{new}$  are much larger than here.

We have contrasted our methods against an upper accuracy bound of 69.1% (model trained on 90 hours of in-domain speech) and a lower bound of 60.8% (no in-domain data). Using only 2 hours of transcribed in-domain speech for in-domain model estimation, we improve over the asymptotic accuracy of the out-of-domain model by 5.1% to 65.9%; two novel ideas, Gaussian sharing and data generation, contribute 2.2% to this improvement. With 9 hours of in-domain speech, the improvement of accuracy is 6.5% (to 67.3%), with a contribution of 1.2% from Gaussian sharing.

We have shown that sharing output densities is an effective method for merging in-domain and out-of-domain models. This merging is most effective when the models are trained on disjoint data sets and have different context dependency structure. Note that the merged models have substantially more parameters resulting in slower decoding. However, the gain in accuracy offsets this for mid to high beams making the technique usable for real time systems.

Further improvements in decoding speed can be achieved using better Gaussian caching and selection.

We enrich the set of transcribed in-domain speech sentences, available for training/adaptation, with artificially generated data that simulates the pronunciation by additional speakers. The models estimated (trained or adapted) on the enriched set are shown more accurate than the models estimated on the original data alone. Our data generation is now based on frequency warping. Future research may explore other production and perception models for data generation.

## 5. References

- [1] M.J.F. Gales, Y. Dong, D. Povey and P.C. Woodland, "Porting: SwitchBoard to the VoiceMail Task", ICASSP 2003, Vol I, 536-540.
- [2] Lori Lamel, Jean-Luc Gauvain, Gilles Adda, "Lightly Supervised Acoustic Model Training", ASR-2000, 150-154.
- [3] H.Y. Chan, P.C. Woodland, "Improving Broadcast News Transcription By Lightly Supervised Discriminative Training", ICASSP-2004.
- [4] Frank Wessel and Herman Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition", ASRU-2001.
- [5] Thomas Kemp and Alex Weibel, "Unsupervised Training Of A Speech Recognizer Using TV Broadcast", ICSLP 98, 2207-2210.
- [6] Young, S.J., Odell, J.J., and Woodland, P.C. "Tree-Based State Tying For High Accuracy Acoustic Modeling", Proc ARPA Human Language Technology Workshop, March 1994, pp. 307-312, Morgan Kaufmann.
- [7] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, pp. 171-185, 1995.
- [8] J.L.Gauvain and C.H.Lee, "Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. Speech and Audio Proc.*, vol 2, pp. 291-298, 1994.
- [9] L.Welling, S.Kanthak, H.Ney, "Improved methods fo Vocal Tract Normalization", *Proc ICASSP 1999*, Vol.2, pp. 761-764.
- [10] X. Luo, *Balancing Model Resolution and Generalizability in Large Vocabulary Continuous Speech Recognition*. PhD Thesis, The Johns Hopkins University, Baltimore, MD, 1999.
- [11] M. Saraclar, H. Nock and S. Khudanpur, "Pronunciation Modeling by Sharing Gaussian Densities Across Phonetic Models," *Computer Speech and Language*, vol 14:2, pp. 137-160, 2000.
- [12] Allen L. Gorin, Alicia Abella, Tirso Alonso, Giuseppe Riccardi, and Jeremy H. Wright, "Automated natural spoken dialogue", *IEEE Computer Magazine*, vol 35, no.4, pp 51-56, 2002.