TASK-SPECIFIC ADAPTATION OF SPEECH RECOGNITION MODELS

Ananth Sankar, Ashvin Kannan, Ben Shahshahani and Eric Jackson[*]

Nuance Communications
1380 Willow Road
Menlo Park, CA 94025

ABSTRACT

Most published adaptation research focuses on speaker adaptation, and on adaptation for noisy channels and background environments. In this paper, we study acoustic, grammar, and combined acoustic and grammar adaptation for creating task-specific recognition models. Comprehensive experimental results are presented using data from a natural language quotes and trading application. The results show that task adaptation gives substantial improvements in both utterance understanding accuracy, and recognition speed.

1. INTRODUCTION

Techniques for acoustic adaptation, such as maximum-likelihood linear regression (MLLR) [1], stochastic matching [2], and maximum a-posteriori (MAP) adaptation [3], have been well studied in the literature. These techniques, and many variations on them have been used for speaker [1], dialect [4], and channel adaptation [2] with considerable success. In this paper, we focus instead on *task adaptation* where general speech recognition models are adapted to a particular application.

Most previous studies have reported only the accuracy improvement from adaptation. However, particularly in real applications, it is important to focus also on speed. We present experimental results to show that acoustic and grammar task adaptation give substantial accuracy and recognition speed improvements over the baseline models. Also, improvements from acoustic and grammar adaptation are shown to be additive.

2. MOTIVATION FOR TASK ADAPTATION

In this paper, we focus on improving the performance of telephone-based speech recognition applications using task-specific adaptation. Examples of such applications include stock-quotes, keywords, digit-strings (such as telephone numbers), and name recognition (for example, in voice dialing).

[*]Now with Google, 2400 Bayshore Parkway, Mountain View, CA 94043

For our experiments, we use task-specific probabilistic finite state grammars (PFSGs), with a dictionary that covers all the words in the test grammars. The hidden Markov model (HMM)-based acoustic models are trained on large amounts of data collected from various applications. These cover various speaking styles, channel conditions, and phonetic contexts.

We achieve very good accuracy using this approach. However, significant improvements can be achieved by adapting the models to each task. Probabilities on grammar paths are clearly task-specific. Learning these probabilities using an automatic data-driven adaptation procedure gives very good results. From an acoustic modeling point of view, one would expect context-dependent phonetic coverage to be task specific. For example, a stock-quote task has different coverage than a digit recognition task. There may even be dialect or speaking style differences specific to each application. Acoustic adaptation can improve performance in these scenarios.

Task adaptation gives sharper models that are better matched to the application data. This results in better accuracy. The adapted models also better discriminate against each other, thus reducing the competing hypotheses during recognition. This results in faster recognition.

3. ADAPTATION ALGORITHMS

For acoustic adaptation, we use a MAP-based smoothing approach motivated by [3]. We adapt the mean (μ) and variance (σ^2) parameters for the HMM Gaussians in the following manner:

$$\begin{aligned} \mu_A &= (1-\lambda)\mu_O + \lambda\mu_D \\ \sigma_A^2 &= (1-\lambda)(\sigma_O^2+\mu_O^2) + \lambda(\sigma_D^2+\mu_D^2) - \mu_A^2, \end{aligned} \quad (1)$$

where

$$\lambda = \frac{N_D}{N_D+\gamma}.$$

The subscript A refers to the adapted parameters, O to the original parameters, and D to the maximum-likelihood (ML) estimates of the parameters computed from the adaptation data. γ is a tuning weight which interpolates between the original model and the adaptation data [5]. A smaller value gives more weight to the adaptation data.

A similar smoothing approach is used for grammar adaptation. The probabilities P_i for each of N possible paths at a point in the grammar are computed by interpolating between the ML estimates computed from the adaptation data, and a uniform distribution:

$$P_{iA} = (1-\lambda)\frac{1}{N} + \lambda\frac{C_i}{\sum_{i=1}^{N} C_i}, \tag{2}$$

where C_i are the number of times path i was observed in the adaptation data, and the weights λ are computed using deleted interpolation.

4. MODES OF ADAPTATION

Task adaptation can be done in two modes: supervised and unsupervised. In supervised adaptation, human transcriptions are used along with the recorded utterances for adaptation. On the other hand, unsupervised adaptation makes use of the recognition hypotheses. Supervised adaptation typically gives better performance than unsupervised adaptation. However, it can only be used in an offline mode. For a fully automatic online approach, unspervised adaptation would be used.

5. USE OF CONFIDENCE SCORES FOR DATA FILTERING

An important aspect of unsupervised adaptation is that the recognition hypotheses are used to adapt the models. These hypotheses could be erroneous, and thus detrimentally affect the adaptation process. If we could automatically select only those utterances whose recognition hypotheses are likely to be correct, adaptation performance could be improved. We use a confidence-score-based approach to address this problem.

Our system computes an integer-valued confidence score between 0 and 100 for each hypothesis based on phone-level posterior probabilities [6]. A score of 0 indicates very low confidence in the recognition result, while a score of 100 indicates very high confidence. In this paper, we use an algorithm that compares these confidence scores against an adaptation confidence threshold. Only utterances that cross the confidence threshold are used for adaptation. We show that this threshold-based filtering gives superior performance for unsupervised adaptation.

6. EXPERIMENTAL RESULTS

The acoustic models we used are based on Genonic hidden Markov models (HMM). In Genonic HMMs [7], triphone states are clustered using bottom-up agglomerative clustering. Each state cluster shares a set of Gaussians (also called a Genone). Each state in a cluster has an independent set of mixture weights to the Gaussians in the shared Genone.

We ran experiments using field data from a natural language stock quotes and trading application. The language for the application is American English. The acoustic models used about 13,000 triphone HMMs, with 1000 Genones, and 32 Gaussians per Genone. We report experimental results on two grammars, denoted as "Main", and "Equities". The "Main" grammar describes various keywords and phrases for navigating the application, and also allows stock quotes, and full natural language trading commands. An example of a trading command is:

> I want to buy five hundred shares of CompanyX
> at twenty dollars

The "Equities" grammar is essentially a stock-quote grammar for a large number of stock symbols. The "Main" test set contains 3253 utterances and the "Equities" test set contains 5141 utterances. Both test sets contained only in-grammar (IG) utterances, i.e., utterances whose reference transcript could be parsed by the respective grammars. In all the experiments below, we use the sentence understanding error-rate on these test sets as our performance measure.

For adaptation, both in-grammar and out-of-grammar (OOG) utterances were used. An adaptation utterance is IG or OOG if its reference transcript (*supervised*), or its recognition hypothesis (*unsupervised*) can or cannot be parsed by the respective grammar. Grammar adaptation was run separately for each of the test grammars, using the IG utterances from the respective grammars. However, acoustic adaptation was run using IG and OOG utterances from all the application grammars. These included the "Main" and "Equities" grammars, and also other grammars in the application. The same adapted acoustic models were used for the two test sets. Real caller data is used; thus the distribution of the data and grammars used in the adaptation set is representative of real system usage.

The baseline system used in this study uses uniform grammar probabilities so as to demonstrate the effect of grammar adaptation. Actual fielded systems for similar applications use grammar probabilities estimated in supervised mode with 200,000 utterances or more, resulting in significantly lower error rates.

6.1. Confidence-Based Data Selection

For unsupervised adaptation, the recognition hypothesis is used to align the data and compute MAP adaptation statistics. To mitigate the effect of recognition errors, we used (for adaptation) only those utterances whose confidence score was above a threshold. Table 1 shows the error rate

Utterance Selection Threshold	Understanding Error (%)	
	Main	Equities
Baseline	15.86	14.76
0	8.82	11.24
30	8.76	11.05
50	8.33	10.46
60	8.67	10.39
70	9.22	10.74

Table 1. Comparison of sentence understanding error-rates for unsupervised ACG adaptation using various confidence thresholds to select 10,000 adaptation utterances

for different confidence thresholds. In each case, 10,000 adaptation utterances were selected. Both acoustic (AC) and grammar (G) models were adapted. We denote this as ACG adaptation.

The table shows that unsupervised adaptation with any threshold is significantly better than the baseline. A threshold of 50 is a good choice. Smaller thresholds result in utterances with incorrect recognition hypotheses being used for adaptation. This results in worse performance for both test sets. Larger thresholds result in the data being biased toward the utterances that were correctly recognized in the first place. This also results in worse performance for both test grammars, though the degradation for the Equities grammar appears to be small. All future unsupervised experiments reported in the paper use a threshold of 50.

6.2. Amount of Adaptation Data

It is expected that increasing the amount of adaptation data would result in a decrease in the error-rate. To evaluate the effect of this, we increased the amount of adaptation data in 5000 utterance chunks. Figure 1 shows the error-rates for supervised ACG adaptation for both test sets. The figure clearly shows that adaptation gives significant improvements for both test sets. The adaptation error clearly decreases with increasing amounts of data. After about 30,000 utterances, the improvements are smaller. However, even at this point, the curves still have a small slope, indicating that using a huge amount of adaptation data ($>> 30,000$ utterances) can give further improvements. While the figure only shows supervised ACG error-rates, we observed similar behavior for all the other adaptation cases.

6.3. Supervised and Unsupervised Adaptation

Since supervised adaptation has access to the correct transcripts, we expect it to perform better than unsupervised adaptation. We compared the performance of supervised and unsupervised ACG adaptation using 35,000 adaptation

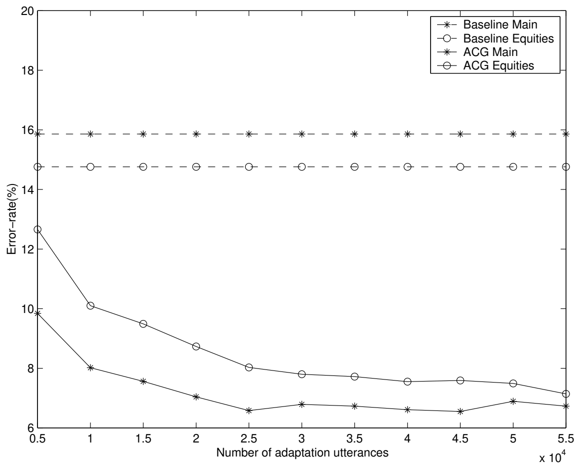

Fig. 1. Effect of varying amount of adaptation data for supervised ACG adaptation

	Main	Equities
Baseline	15.86	14.76
Unsup ACG (Improvement(%))	7.62 (52.0)	8.60 (41.7)
Sup ACG (Improvement(%))	6.73 (57.6)	7.72 (47.7)

Table 2. Sentence understanding error rates for supervised and unsupervised ACG adaptation using 35,000 adaptation utterances. Percentage relative improvements over baseline in parenthesis

utterances. The results are given in Table 2. Both supervised (Sup) and unsupervised (Unsup) adaptation give a huge gain over the baseline. Further, supervised adaptation gives a significant improvement over unsupervised adaptation (about 10% relative) for both test grammars.

6.4. Effect of Acoustic and Grammar Adaptation

Since acoustic and grammar adaptation modify different recognition models, the improvements from both are likely to be additive. We use supervised adaptation on 35,000 utterances to compare the effect of acoustic and grammar adaptation on both the test sets. Table 3 shows the error-rate for acoustic only (AC), grammar only (G), and combined acoustic and grammar (ACG) supervised adaptation. It is clear that grammar adaptation gives a bigger improvement than acoustic adaptation for this task. For example, on the "Main" grammar test set, grammar adaptation gave a 52.0% improvement, while acoustic adaptation gave a 11.2% improvement. However, the improvements are nearly additive, so that combined acoustic and grammar adaptation gave a 57.6% improvement.

	Main	Equities
Baseline	15.86	14.76
Sup AC (Improvement(%))	14.08 (11.2)	13.01 (11.9)
Sup G (Improvement(%))	7.62 (52.0)	8.15 (44.8)
Sup ACG (Improvement(%))	6.73 (57.6)	7.72 (47.7)

Table 3. Sentence understanding error rates for supervised acoustic, grammar, and combined acoustic and grammar adaptation using 35,000 utterances. Percentage improvements over baseline in parenthesis

6.5. Effect of Adaptation on Recognition Speed

Adaptation results in models that are better matched to the test condition. The resulting models are sharper and hence have higher discrimination against competing models. This results in fewer hypotheses being maintained in the Viterbi beam search, giving faster recognition. Recognition accuracy can be traded off for speed. Thus, it is important to jointly view accuracy and speed, especially for real applications.

Figure 2 shows the speed-accuracy trade-offs for the baseline models and different adapted models for the Equities test set. The curves in the figure are for the case of unsupervised adaptation using 35000 utterances. The points on the curve represent different pruning values in the Viterbi beam search. The large circles on each curve represent the operating point using the default pruning value used by the actual application. In the figure, the recognition times are given as a factor of the baseline system's actual recognition time using the application's default pruning value. If we maintain the error-rate at the baseline level, then the figure shows that acoustic (AC) adaptation gives a 25% speedup, grammar (G) adaptation gives a 45% speedup, and combined acoustic and grammar adaptation (ACG) gives a 50% speedup. As the large circles in the figure show, adapted models are also faster at the same pruning threshold. Thus, if we maintain the original baseline recognition speed by increasing the pruning value for the adapted models, we can get an even bigger accuracy improvement than the substantial ones already shown in previous sections.

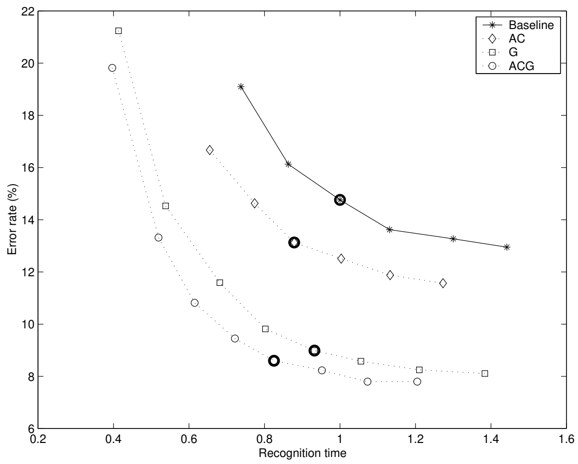

Fig. 2. Effect of unsupervised adaptation with 35,000 utterances on speed and accuracy

7. SUMMARY AND CONCLUSIONS

This paper presented a detailed study of task adaptation. Both supervised and unsupervised adaptation was studied, as was acoustic, grammar, and combined adaptation. We presented a confidence-score-based filtering algorithm to address the problem of erroneous recognition hypotheses for unsupervised adaptation. Detailed experimental studies indicate that the approach gives substantial improvements, both in accuracy and speed, over the baseline models.

8. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," *Proc. Spoken Lang. Systems Technology Workshop*, pp. 110–115, 1995.

[2] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE TSAP*, vol. 4, pp. 190–202, May 1996.

[3] J. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE TSAP*, vol. 2, pp. 291–298, April 1994.

[4] V. Diakoloukas, V. Digalakis, L. Neumeyer, and J. Kaja, "Development of Dialect-Specific Speech Recognizers Using Adaptation Methods," *Proc. ICASSP*, vol. 2, pp. 1455–1258, 1997.

[5] D. A. Reynolds and T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] E. Chang, "Improving Rejection with Semantic Slot-Based Confidence Scores," in *Proc. EUROSPEECH*, pp. 271–274, 1999.

[7] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE TSAP*, vol. 4, no. 4, pp. 281–289, 1996.