USING A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNIZER FOR A CONSTRAINED DOMAIN WITH LIMITED TRAINING

Manhung Siu, Michael Jonas, Herbert Gish

BBN Technologies/GTE Internetworking,
70 Fawcett Str., Cambridge, MA 02138
msiu@bbn.com

ABSTRACT

How to train a speech recognizer with limited amount of training data is of interest to many researcher. In this paper, we describe how we use BBN's Byblos large vocabulary continuous speech recognition (LVCSR) system for the military air-traffic-control domain where we have less than an hour of training data. We investigate three ways to deal with the limited training data: 1) re-configure the LVCSR system to use fewer parameters, 2) incorporate out-of-domain data, and, 3) use pragmatic information, such as speaker identity and controller function to improve recognition performance. We compare the LVCSR performance to that of the tied-mixture recognizer that is designed for limited vocabulary. We show that the reconfigured LVCSR system outperforms the tied-mixture system by 10% in absolute word error rate. When enough data is available per speaker, vocal tract length normalization and supervised adaptation techniques can further improve performance by 6% even for this domain with limited training. We also show that the use of out-of-domain data and pragmatic information, if available, can each further improve performance by 1-3%.

1. INTRODUCTION

In our previous papers [1] [2], we have reported speech recognition results on the air-traffic control (ATC) domain using the tied-mixture recognizer that is designed for limited vocabulary. In recent years, research in large vocabulary continuous speech recognition (LVCSR) has made tremendous progress. Many sites have been reporting more than 10% relative error rate reduction each year in tasks such as Switchboard and Wall Street Journal transcription. Most of the recognition systems use millions of parameters that can only be trained with tens to hundreds of hours of data in the desired domain. In the ATC domain, the amount of training data is orders of magnitude less. The question is then how to use an LVCSR system for a domain with limited training data.

In this paper, we describe three approaches to use BBN's Byblos LVCSR system for the military air-traffic (MATC) domain where we have less than an hour of training data. First, we show how to re-configure the LVCSR system to use fewer parameters, and examine the effectiveness of two techniques that have been very successful in LVCSR: vocal tract length normalization and unsupervised speaker adaptation. Second, we incorporate out-of-domain data to augment the limited training. We introduce the use of supervised adaptation technique to incorporate out-of-domain acoustic training data. Third, we incorporate pragmatic information, such as speaker identity and controller function that are available to the MATC domain. We compare the LVCSR performance to that of a tied-mixture recognizer that is designed for limited vocabulary and find the re-configured LVCSR engine to be superior. When enough training is available per speaker, both the vocal tract length normalization and the unsupervised adaptation techniques are effective. We also show that the use of out-of-domain data and pragmatic information, if available can further enhance performance.

The remaining portion of the paper is organized as follows. In Section 2, we describe the Greenflag corpus and compare its characteristics with the Switchboard Corpus on which Byblos is developed. In Section 3, we describe our approaches to deal with limited data, including the reconfiguration of Byblos, incorporation of out-of-domain data and incorporation of pragmatic information. In Section 4, we describe our experiments and we conclude with a discussion in Section 5.

2. MILITARY AIR TRAFFIC CONTROL AND THE GREENFLAG CORPUS

The Military Air Traffic Control (MATC) domain consists of off-the-air recorded conversations between military pilots and controllers. Similar to the civilian ATC, these recordings consist of sequences of interleaving exchanges between different pilots and a single controller. The length of these exchanges can vary from several words to several tens of words. The language is highly constrained but the acoustic quality is quite poor. The language and acoustic quality differ between pilots and controllers. While in civilian ATC, pilots converse only to controllers and never among themselves, this is not the case in MATC where pilots within a flight group sometimes talk to each other.

The Greenflag (GNFG) Corpus is made up of a number of recorded sessions of U.S air force exercises. Each session

contains about 5 minutes of speech. In addition, different controllers are marked for their functions, such as approach, tower or ground. In Table 1, we summarize the data characteristics of the GNFG Corpus as compared to the Logan Corpus (LG) of civilian ATC and the Switchboard (SWBD) corpus of telephone conversational speech. We notice that the amount of data per speaker is limited for GNFG pilots. The signal to noise ratio (SNR) reported in Table 1 is measured by comparing the 65-th energy percentile to the 5-th percentile. The SNR of GNFG pilots is significantly worse than any other corpora. Perplexity is measured using a trigram language model on our held-out test-set. The perplexity of the GNFG corpus is significantly lower than other corpora, reflecting the highly structured language. The oov % denotes the rate of out-of-vocabulary (oov) as measured by the proportion of word tokens in our test-set that is not observed in training. The oov rate is significantly higher in GNFG partly because of the limited amount of training data. It is also caused by the fact that military flight call-signs are made up of randomly selected flight names such as "echo" or "viper" which rarely overlap between training and test.

3. HANDLING LIMITED TRAINING DATA

We propose three different approaches to handle limited training data. First, we reconfigure the LVCSR engine to reduce the number of parameters. Second, we incorporate out-of-domain acoustic and language model training data. Third, we try to capitalize pragmatic information that is available in the MATC domain.

3.1. Re-configure Byblos for limited data

Byblos is a continuous density HMM-based recognizer [3] that uses phonetic models to represent its vocabulary. Each phone is represented by a 5-state HMM and its state observation distribution is represented by Gaussian mixtures. Byblos maintains two types of phonetic model, the phonetic-tied-mixture (PTM) model and the State-clustered-tied mixture (SCTM) model. They differ by how they tie model parameters. The coarser PTM model uses a single set of Gaussians for each context-independent phone to represent the observation distributions. This set of Gaussians is shared across all context-dependent phones of the same center phone. Separate mixture weights are estimated for different states of different context-dependent phones. The more detailed SCTM model uses a more general sharing strategy and allows states of the same context-independent phones to use different set of Gaussians. Search in Byblos is done in multiple passes. In the first few passes, the coarser PTM non-crossword models are used in conjunction with a bigram language model. In the last pass, the more detailed SCTM crossword models are used in conjunction with a trigram language model.

The limited amount of training in the GNFG Corpus require us to reduce the number of parameters used in Byblos. Of the two types of model in Byblos, the SCTM model requires a lot more parameters. Even with the built-in back-off in state-clustering, the amount of training by GNFG is far too limited for SCTM. Instead, we use only the PTM model. One variable that controls the number of parameters in the PTM model is the number of Gaussian mixtures. For a PTM model with K Gaussians per phone, the number of parameters is $2LKD + NK$, where L is the number of context independent phones, N is the number of context-dependent phones with unique mixture weights and D is the number of features. Thus, the number of Gaussians to use, while dependent on the amount of training data, is also dependent on the number of context-dependent phones N. We determine the best K for the GNFG corpus experimentally.

VTL normalization [4] and unsupervised speaker adaptation [5] are very successful techniques in the SWBD transcription task where each has shown to reduce the absolute word error rate by approximately 4%. Both techniques aim at reducing the variability between speakers by estimating some compensation factors for each speaker. In VTL, the compensation factor is a single frequency warp per speaker. In unsupervised adaptation, the compensation factors are a number of linear transformations. Being able to robustly estimate these compensation factors is the key to the success of these techniques and is dependent on the amount of data per speaker. Furthermore, since VTL stretch is estimated at the front-end, it may be more sensitive to the SNR. The unsupervised adaptation on the other hand, relies on the errorful transcription from the recognizer and its performance may be dependent on the recognition accuracy.

In SWBD, the amount of data per speaker (conversation side) is around 2.5 minutes. In GNFG, the condition differs dramatically between pilots and controllers. For controllers, the amount of data per speaker is comparable to SWBD while pilots have very limited data per speaker. Controllers also have better SNR than pilots. In Section 4, we report results of using VTL and unsupervised speaker adaptation on GNFG pilots and controllers.

3.2. Using Out-of-domain Data

One approach to deal with limited training is to incorporate training data from out-of-domain. However, pooling the data together can dilute the contributions of in-domain training. For acoustic modeling, instead of pooling the in-domain and out-of-domain data, we use supervised speaker adaptation to perform domain adaptation. First, we train our acoustic model using out-of-domain data from domains such as LG or SWBD. This forms our basic acoustic model. Then we use our in-domain training data and its training transcription to estimate adaptation transformations that shift the out-of-domain model to the space of the in-domain data.

For language modeling, we use a weighted combination scheme that was first proposed in [6]. A metric measuring the similarity between the out-of-domain data and the in-

Corpus	Corpus size (hr)	Data/spkr (sec)	Sent. len. (sec)	SNR	Data type	Perpl.	Voca. size	OOV (%)
SWBD	160	150	2.0	22	telephone	100	25K+	1
LG ctrl.	2.5	730	3.8	15	radio	20	800	1
LG pilot.	1.8	14	2.6	16	radio	39	650	12
GNFG ctrl.	0.75	142	2.8	21	radio	19	600	13
GNFG pilot.	0.5	12	1.6	12	radio	20	500	10

Table 1: Data characteristics of GNFG, LG and SWBD

domain training is estimated on each piece of out-of-domain data. This similarity is then used to weight the relative importance of the out-of-domain data when combining with the in-domain data.

Our techniques of combining out-of-domain data implicitly assume that some similarity between the out-of-domain data and in-domain data exist such that the transformations or weighting can steer the model to the right place.

3.3. Using Pragmatic Information

In this section, we consider two pieces of pragmatic information, the speaker identity and the controller function. From a practical point of view, there are only a limited number of controller in the tower and knowing which one it is can be instrumental in improving the recognition performance. Similar argument can be made on the controller functions.

Suppose a test speaker is identified and is one of the training speakers. What is the best way to use this information? One approach is to use a speaker dependent model. Given the limited amount of data we have in GNFG, building a speaker dependent model is impossible. We take the approach of using supervised speaker adaptation to transform the speaker independent model closer to the test speaker.

When the the controller function is known, we can use this information in building a function-specific language model. However, a function-specific model, similar to speaker dependent model in acoustic modeling, causes fragmentation of data. Instead, we use an approach similar to the use of out-of-domain data. We consider the function-specific data as in-domain and the general data as out-of-domain and apply the exact algorithm as describe above for incorporating out-of-domain language modeling data.

4. EXPERIMENTS

4.1. Paradigms

Our GNFG training data consists of a total of 35 minutes of controller and 24 minutes of pilot speech. The test data consists of 10 minutes of controller and 6 minutes of pilot speech. There are a total of 14 different controllers and 130 different pilots in training and 10 different controllers and 71 different pilots in test. Separate models are built for pilot and controllers because of their differences in acoustic and language characteristics. Test and training come from different recording sessions. However, 3 speakers out of the 10 speakers in the controller test is also in training. For experiments using out-of-domain data, a total of 131 minutes of Logan ATC controller data is used consisting of 10 speakers.

Except for the contrast experiment using the tied-mixture recognizer, all recognition experiments use the Byblos recognizer. A total of 45 features are used, including the normalized energy, 14 cepstral coefficients, and their first and second order differences. A PTM cross-word model is used in conjunction with a trigram language model. VTL normalization is applied per speaker for both pilots and controllers. Results for all recognition experiments are reported in terms of word error rates.

4.2. Experiments

We performed two sets of experiments. In the first set, we test the baseline Byblos system by selecting the best number of Gaussians in the PTM model. Then, we test the effectiveness of unsupervised adaptation and VTL normalization. In the second and third set, we tested the use of out-of-domain data and use of pragmatic information for the controllers only, where VTL normalization is applied but unsupervised speaker adaptation is not.

Five different PTM model sizes are tested as shown in Table 2. We notice that the best size for both pilots and controllers are 64 Gaussians mixtures. It should be noted that the fewer the number of Gaussians, the faster the recognition speed since a significant portion of recognition processing time is spent on evaluating the Gaussian mixtures. Thus, while we use 64 Gaussians for in our experiments, the 32 Gaussian mixtures is a good choice for a faster implementation.

We test the effectiveness of VTL normalization for both pilot and controllers and the results are shown in Table 3. The controllers improves by 2.5% while the pilots degrade by 8.5%. We also test the effectiveness of unsupervised adaptation for both pilot and controllers with VTL and the results are shown in Table 4. We notice that unsupervised adaptation improves recognition performance of controllers by about 4% absolute which is consistent with the gain in the SWBD corpus. However, it helps pilots only by 1%. Comparing the amount of data per speaker as tabulated in Table 1, we notice that the controllers have around 2.5 minutes on average per speaker which is very similar to that of the SWBD speakers. The pilots, however, have only 14 seconds of training and its SNR is 12 compare to 20 of the

Speakers	number of mixtures				
	256	128	64	32	16
Controllers	39.8	38.6	36.7	37.8	40.4
Pilots	52.0	51.9	49.7	52.2	55.0

Table 2: Finding the best number of Gaussian mixtures per phone

Speakers	with VTL	without VTL
Controllers	36.7	39.3
Pilots	49.7	41.1

Table 3: Effect of VTL on pilots and controllers

controllers.

We can also compare the Byblos result with those of the tied-mixture system [1] that is trained under comparable condition. On a slightly smaller test-set on controllers without VTL normalization nor adaptation, the tied-mixture system gives a recognition error of 47% while the Byblos system gives a recognition error of about 36% showing that the use of the LVCSR system can be re-configured to outperform a limited vocabulary system.

We tested the effect of adding LG data for both acoustic modeling (AM) and language modeling (LM) as tabulated in Table 5. The experiments reported are on controllers only using VTL with a GNFG trained trigram language model. We notice that the use of supervised adaptation improves recognition by 1.4% while the adaptation of language modeling data hurts slightly. We also report the result of using the LG acoustic model and SWBD acoustic models as is. While using LG only or SWBD only is significantly worse (8%) than using GNFG, it is interesting to see the use of SWBD with 120 hrs of training of very different channel is comparable to the use of LG with 2.5 hrs of similar channel. One possible explanation is that the degradation due to channel mismatch is compensated by an increase in training. On the other hand, the addition of LG data for language modeling is not useful. It is not surprising given that the use of LG language model doubles the word error rates indicating that the language is quite different between GNFG and LG.

We also tested the effect of using supervised speaker adaptation on 3 GNFG controllers and the effect of adapting the tower controller function. The results are tabulated in Table 6. We notice that the supervised adaptation improve recognition by about 3%. It should be noted that the speakers' data are already used in training the model before adaptation. For the activity adaptation on language model, we obtained a 0.5% gain. Similar to that of the supervised

Speakers	Unadapted	Adapted
Controllers	36.7	33.2
Pilots	41.1	40.3

Table 4: Effect of adaptation on pilots and controllers

Expts.	GNFG	LG	SWBD	LG + GNFG
AM	36.7	46.3	45.2	35.3
LM	36.7	77.0	–	37.0

Table 5: Effect of using out-of-domain data in word error rate

Experiments	Before Adaptation	Adapted
Acoustic Model	29.4	26.4
Language Model	31.1	30.5

Table 6: The effect of pragmatic information on recognition word error rate

adaptation result, training for this activity is already part of the original training before adaptation. For a test set of this size, this gain may not be statistically significant.

5. DISCUSSION

In this paper, we showed that the by using PTM model with 64 Gaussian components, the Byblos LVCSR system outperforms the tied-mixture system by approximately 10%. VTL normalization and unsupervised speaker adaptation further improve the controller performance by 6.5%. These techniques, however, are not as useful on pilots probably due to insufficient data per speaker and poor acoustic quality. We further show that supervised adaptation technique can be used to transform models trained for a different domain to the target domain. Knowing the identity of a test controller speaker who is also in training can further improve the recognition performance by 3% and knowing the controller function is marginally useful.

6. ACKNOWLEDGEMENT

This work was supported by Air Force Research Labs under contract number 1F30602-97-C-0048

7. REFERENCE

1. J.R. Rohlicek *et al.*, "Gisting conversational speech", *ICASSP*, pp. II:113-116, 1992.
2. L. Denenberg *et al.*, "Gisting conversational speech in real time", *ICASSP*, pp. II:131-134, 1993.
3. J. Billa *et al.*, "Multi-lingual speech recognition: The 1996 BYBLOS Callhome system," in *Proc. EuroSpeech*, pp. 363–366, 1997.
4. S. Wegmann *et al.*, "Speaker normalization on conversational telephone speech," in *ICASSP*, pp. 339–341, 1996.
5. C. Leggetter and P. Woodland "Maximum likelihood linear regression for speaker adaptation of HMMs," in *Computer Speech and Language*, pp. 171-186, 1995.
6. R. Iyer, M. Ostendorf and H. Gish, "Using Out-of-Domain Data to Improve In-Domain Language Models," *IEEE Signal Processing Letters*, vol. 4, no. 8, pp. 221-223, August 1997