Accent Adaptation for the Air Traffic Control Domain

Matthew Garber, Meital Singer, Christopher Ward
Brandeis University
Waltham, MA 02453, USA
{mgarber,tsinger,cmward}@brandeis.edu

Abstract

Automated speech recognition (ASR) plays a significant role in training and simulation systems for air traffic controllers. However, because English is the default language used in air traffic control (ATC), ASR systems often encounter difficulty with speakers' non-native accents, for which there is a paucity of data. This paper examines the effects of accent adaptation on the recognition of non-native English speech in the ATC domain. Accent adaptation has been demonstrated to be an effective way to model under-resourced speech, and can be applied to a variety of models. We use Subspace Gaussian Mixture Models (SGMMs) with the Kaldi Speech Recognition Toolkit to adapt acoustic models from American English to German-accented English, and compare it against other adaptation methods. Our results provide additional evidence that SGMMs can be an efficient and effective way to approach this problem, particularly with smaller amounts of accented training data.

1 Introduction

As the field of speech recognition has developed, ASR systems have grown increasingly useful for the ATC domain. The majority of air traffic communication is verbal (Hofbauer et al., 2008), meaning ASR has the potential to be an invaluable tool not just in assisting air traffic controllers in their daily operations, but also for training purposes and workload analysis (Cordero et al., 2012).

Due to a constrained grammar and vocabulary, ATC ASR systems have relatively low word error rates (WER) when compared to other domains, such as broadcast news (Geacăr, 2010). These systems can also be limited at run-time by location (e.g. place names, runway designations), further constraining these parameters and increasing accuracy.

Despite the effectiveness of existing systems, air traffic control has little tolerance for mistakes in day-to-day operations (Hofbauer et al., 2008). Furthermore, these systems generally perform worse in real-world conditions, where they have to contend with confounding factors such as noise and speaker accents (Geacăr, 2010).

In this paper, we attempt to ameliorate the issue of speaker accents by examining the usefulness of accent adaptation in the ATC domain. We compare the relatively new innovation of SGMMs (Povey et al., 2011a) against older adaptation techniques, such as maximum a posteriori (MAP) estimation, as well as pooling, a type of multi-condition training.

We perform experiments using out-of-domain American English data from the HUB4 Broadcast News Corpus (Fiscus et al., 1998), as well as German-accented English data taken from the ATCOSIM corpus (Hofbauer et al., 2008) and provided by UFA, Inc., a company specializing in ATC training and simulation.

The paper is organized as follows: in Section 2, we describe previous accent adaptation techniques as well as the structure of SGMMs and how they can be adapted on new data. In Section 3, we outline our experiments and show how accent adaptation with SGMMs outperforms other methods when using smaller amounts of data. Section 4 concludes the paper and presents paths for future study.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 95–99
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-3016

2 Background

2.1 Accent Adaptation

The ideal ASR system for non-native accented speech is one trained on many hours of speech in the target accent. However, for a variety of reasons, there is often a paucity of such data. Several different techniques have been employed to model accented speech in spite of this lack of data.

One method is to manually adjust the pronunciation lexicon to match the accented phone set (Humphries et al., 1996). Unfortunately, this is both time- and labor-intensive as it requires mappings to be generated from one phoneset to another, either probabilistically or using expert knowledge.

Another technique is interpolate models, with one trained on native accented speech and the other trained on non-native accented speech (Witt and Young, 1999). While this has been shown to reduce word error rate (Wang et al., 2003), it does not fully adapt the native model to the new accent.

An effective and versatile method, and the one we implement here, is to directly adapt a native acoustic model on the non-native speech. There exist a few different ways to accomplish this, such as MAP estimation for HMM-GMMs, Maximum Likelihood Linear Regression (MLLR) for Gaussian parameters (Witt and Young, 1999), and re-estimating the state-specific parameters of an SGMM. These techniques have the advantage of requiring little other than a trained native accent model and a non-trivial amount of non-native accented data.

2.2 SGMM Adaptation

Unlike a typical GMM, the parameters of an SGMM are determined by a combination of globally-shared parameters and state-specific parameters. The model can be expressed as follows:

$$p(\mathbf{x}|j) = \Sigma_{m=1}^{M_j} c_{jm} \Sigma_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \mathbf{\Sigma}_i)$$

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\Sigma_{l=1}^{I} \exp \mathbf{w}_l^T \mathbf{v}_{jm}}$$

where $\mathbf{x}$ is the feature vector, j is the speech state, c_{jm} is the state-specific weight, $\mathbf{v}_{jm}$ is the state-specific vector, and $\boldsymbol{\mu}_{jmi}$, $\mathbf{M}_i$, and $\mathbf{w}_i$ are all globally-shared parameters. SGMMs are generally initialized using a Universal Background Model (UBM), which is trained separately from the SGMM.

SGMMs can be further extended beyond the model described above to include speaker-specific vectors and projections. Other speaker adaptation techniques, such as feature-space Maximum Likelihood Linear Regression (fMLLR, also known as CMLLR), can be applied on top of these extensions to further increase the accuracy of the model.

Though it is possible to perform MAP adaptation using SGMMs, their unique structure allows a different and more effective technique to be applied (Povey et al., 2011a). Initially, all of the model's state-specific and globally-shared parameters are trained on out-of-domain data. The state-specific parameter $\mathbf{v}_{jm}$ can then be re-estimated on the non-native speech using maximum likelihood. This adaptation method has been successfully applied to multi-lingual SGMMs (Povey et al., 2011a) as well as different native accents of the same language (Motlicek et al., 2013), and is the technique we use here to adapt the native-accented SGMMs on non-native speech.

Though there exist other ways of adapting SGMMs (Juan et al. (2015) created a multi-accent SGMM by combining UBMs that had been separately trained on the native and non-native data), we do not implement those methods here.

3 Experiments

All experiments were performed using the Kaldi Speech Recognition Toolkit (Povey et al., 2011b).

3.1 Data

Speech Data

For acoustic model training, data was taken from three separate sources:

- Approximately 75 hours of US English audio was taken from the 1997 English Broadcast News Corpus (HUB4), which consists of various radio and television news broadcasts.

- About 20 hours of German-accented data, which is purely in-domain ATC speech, was supplied by UFA.

- An additional 6 hours of German-accented speech was taken from the ATCOSIM corpus, which consists of audio recorded during real-time ATC simulations.

	Native (Unadapted)		Non-native Only		Pooled (Unadapted)		Native (Adapted)		Pooled (Adapted)	
System	WER	SER	WER	SER	WER	SER	WER	SER	WER	SER
HMM-GMM	25.73	74.39	5.74	34.43	6.76	39.24	6.55	37.43	5.38	32.77
+ fMLLR	12.61	58.40	4.64	30.64	5.40	34.53	5.12	33.17	4.51	29.30
SGMM	13.78	58.43	4.15	26.23	4.97	30.99	4.13	26.65	3.71	24.44
+ fMLLR	10.02	51.76	3.46	23.57	4.38	29.15	4.09	27.57	3.25	22.38

Table 1: Error rates of different models trained with 6.5 hours of adaptation data.

	Native (Unadapted)		Non-native Only		Pooled (Unadapted)		Native (Adapted)		Pooled (Adapted)	
System	WER	SER	WER	SER	WER	SER	WER	SER	WER	SER
HMM-GMM	25.73	74.39	3.90	25.16	4.99	31.11	5.95	35.18	4.29	27.54
+ fMLLR	12.61	58.40	3.36	23.00	4.16	27.84	4.79	31.36	3.66	24.91
SGMM	13.78	58.43	3.12	21.34	3.82	25.29	3.71	24.71	3.00	21.10
+ fMLLR	10.02	51.76	2.77	20.03	3.34	23.18	3.64	25.16	2.88	20.45

Table 2: Error rates of different models trained with 26 hours of adaptation data.

Test data consisted of just under 4 hours of German-accented speech provided by UFA. All audio was downsampled to 16 kHz.

Language Model Data

We interpolated a language model trained on the UFA and ATCOSIM utterances with one trained on sentences generated from an ATC grammar supplied by UFA. Both LMs were 5-gram models with Witten-Bell smoothing. The interpolated model was heavily weighted towards the natural utterances (with $\lambda = 0.95$), since the main purpose of the generated utterances was to add coverage for words and n-grams that were not present in the natural data.

Lexicon

The lexicon was largely derived from the CMU Pronouncing Dictionary. Additional pronunciations were supplied by UFA, and several were written by hand.

3.2 Experimental Setup

The baseline acoustic model was a regular HMM-GMM and was trained with the usual 39 MFCC features, including delta and acceleration features. Experiments were conducted both with and without pooling the adaptation data with the US English data, since pooling data prior to adaptation has been shown to give better results for both MAP and SGMM maximum likelihood adaptation (Motlicek et al., 2013), as well as for other adaptation techniques (Witt and Young, 1999).

We conducted two experiments, each with a different amount of adaptation data. The first experiment included only a 6.5-hour subset of the total adaptation data, which was created by randomly selecting speakers from both the ATCOSIM corpus and the UFA data. The second included all 26 hours of adaptation data. HMM-GMM models were adapted using MAP estimation and SGMMs were adapted using the method outlined above.

For each amount of adaptation data, we trained several different models, testing all combinations of the following variables:

- Whether the model was trained solely on the native-accented data, trained solely on the adaptation data, trained on the combined data but not adapted, trained of the native data and then adapted, or trained on the combined data and then adapted.
- Whether an HMM-GMM or SGMM was used (as well as the corresponding adaptation method).
- Whether the model was trained with speaker-dependent fMLLR transforms.

3.3 Experimental Results

With 6.5 hours of German accented data, both the MAP-adapted and SGMM-adapted pooled systems saw modest reductions in word error rate, as can be seen in Table 1. MAP adaptation provided a 6.3% relative improvement over the corresponding accented-only model[1], though WER was reduced by only 2.8% when fMLLR was implemented. Pooled SGMMs were more versatile and amenable to adaptation, with a relative reduction in WER of 10.6%, and a relative improvement of

[1] Unless otherwise specified, reduction in error rate is relative to the model with the same parameters (SGMM, fMLLR, etc.) but trained on the non-native accented data only, rather than relative to a single baseline.

6.1% when using fMLLR. Changes in sentence error rate (SER) between models correlated with the changes in WER, reaching a minimum of 22.38% with the adapted SGMM-fMLLR system, a relative reduction of just over 5%.

Including the full 26 hours of non-native speech in the training and adaptation data generally resulted in higher error rates in the adapted systems than the corresponding accented-only models, as seen in Table 2. This decrease in performance approached 10% for the HMM-GMM systems. Though the SGMM-fMLLR adapted system experienced a relative reduction in performance of about 4%, the performance of the non-fMLLR SGMM increased by about the same amount. Changes in SER again correlated with the changes in WER, with the adapted speaker independent SGMM possessing a slight edge (about 1%) over its accented-only counterpart.

It is not clear from this experiment why the speaker independent SGMM system was the only one to undergo an increase in performance when adapted with the full dataset. A possible explanation is that, with enough data, the speaker adaptive techniques were simply more robust than the accent-adaptation method.

Unsurprisingly, the unadapted native-accented systems had the worst performance out of all of the models, with word error rates that were more than double than that of next best corresponding system.

The unadapted pooled models and the adapted native models were usually the second- and third-worst performing groups of models, though their ranking depended on the amount of adaptation data used. The pooled models generally gave better results when more adaptation data was provided, while the adapted native models had an advantage with less adaptation data.

Interestingly, fMLLR had relatively little effect when used with the adapted native SGMMs, regardless of the amount of adaptation data used. WER was reduced by only about 1 to 2% compared to the models' non-fMLLR counterparts. This stands in contrast with the gains that virtually every other model saw with the introduction of fMLLR. It is not clear why this was the case, though it might relate to some overlap between the SGMM adaptation method and fMLLR.

While it is possible that training the pooled model with in-domain English speech could increase performance, it seems unlikely that it would be superior to either the accented-only model or the adapted pooled model.

4 Conclusion

In this paper, we explored how non-native accent adaptation can be applied using SGMMs to yield notable improvements over the baseline model, particularly when there exists only limited in-domain data. We also demonstrated that this technique can achieve as high as a 10% relative improvement in WER in the ATC domain, where the baseline model is already highly accurate. Even with large amounts of adaptation data, speaker independent SGMMs saw a minor increase in performance when adapted, compared to when they were trained only with in-domain data.

Future avenues of research include whether the SGMM adaptation technique used here could be successfully combined with the UBM-focused adaptation method used by Juan et al. (2015) to achieve even further reductions in WER.

Furthermore, future work could explore whether smaller error rates could be achieved by training the original acoustic models on speech from the ATC domain, rather than from broadcast news, and whether the increases in perfomance found here still hold between more distantly related and phonologically dissimilar languages. It should be noted, however, that this may necesitate the creation of new corpora, as the few non-native ATC corpora that exist seem to only include European accents.

References

José Manuel Cordero, Manuel Dorado, and José Miguel de Pablo. 2012. Automated speech recognition in atc environment. In *Proceedings of ATACCS 2012, the 2nd International Conference on Application and Theory of Automation in Command and Control Systems*. IRIT Press, pages 46–53.

Jonathan Fiscus, John Garofolo, Mark Przybocki, William Fisher, and David Pallett. 1998. *1997 English Broadcast News Speech (HUB4) LDC98S71*. Linguistic Data Consortium, Philadelphia.

Claudiu-Mihai Geacăr. 2010. Reducing pilot/atc communication errors using voice recognition. In *Proceedings of ICAS 2010, the 27th Congress of the Internation Council of the Aeronautical Sciences*.

Konrad Hofbauer, Stefan Petrik, and Horst Hering. 2008. The atcosim corpus of non-prompted clean air traffic control speech. In *Proceedings of LREC 2008, the Sixth International Conference on Language Resources and Evaluation*.

Jason J. Humphries, Philip C. Woodland, and D. Pearce. 1996. Using accent-specific pronunciation modelling for robust speech recognition. In *Proceedings of ICSPL 96, Fourth International Conference on Spoken Language, 1996*. IEEE, volume 4, pages 2324–2327.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Tien-Ping Tan. 2015. Merging of native and non-native speech for low-resource accented asr. In *Proceedings of SLSP 2015, the Third International Conference on Statistical Language and Speech Processing*. Springer, pages 255–266.

Petr Motlicek, Philip N. Garner, Namhoon Kim, and Jeongmi Cho. 2013. Accent adaptation using subspace gaussian mixture models. In *Proceedings of ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pages 7170–7174.

Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, et al. 2011a. The subspace gaussian mixture modela structured model for speech recognition. *Computer Speech & Language* 25(2):404–439.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011b. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, EPFL-CONF-192584.

Zhirong Wang, Tanja Schultz, and Alex Waibel. 2003. Comparison of acoustic model adaptation techniques on non-native speech. In *Proceedings of ICASSP 2003, IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, volume 1, pages 540–543.

Silke M. Witt and Steve J. Young. 1999. Off-line acoustic modelling of non-native accents. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH 1999)*. pages 1367–1370.