# Automatic Speech Recognition of Multiple Accented English Data

*Dimitra Vergyri*[1] [*], *Lori Lamel*[2], *Jean-Luc Gauvain*[2] [†]

[1]SRI International, Speech Technology and Research Lab, Menlo Park, CA
[2]LIMSI-CNRS, Spoken Language Processing Group 91403 Orsay cedex, France

dverg@speech.sri.com, {lamel,gauvain}@limsi.fr

## Abstract

Accent variability is an important factor in speech that can significantly degrade automatic speech recognition performance. We investigate the effect of multiple accents on an English broadcast news recognition system. A multi-accented English corpus is used for the task, including broadcast news segments from 6 different geographic regions: US, Great Britain, Australia, North Africa, Middle East and India. There is significant performance degradation of a baseline system trained on only US data when confronted with shows from other regions. The results improve significantly when data from all the regions are included for accent-independent acoustic model training. Further improvements are achieved when MAP-adapted accent-dependent models are used in conjunction with a GMM accent classifier.

**Index Terms**: accented speech recognition, accent adaptation

## 1. Introduction

Speaker variability, such as gender, accent, age, speaking rate, and phone realizations, is an important difficulty in automatic speech recognition, affecting the performance as much as noise and channel variability. Any deployed speech recognition system should exhibit robustness in such variability in order to be useful. Despite large progress in large vocabulary speech recognition in the fields of noise and channel robustness, speaker normalization for age and gender, and unsupervised speaker adaptation that compensates for some of the speaker variability, recognition accuracy has been observed to drastically degrade when the accent of the speaker deviates from the standard accent in the training data, as in the case for non-native speakers of the target language [1] or speakers with regional accent not present in the acoustic training data [2].

Recent work has focused on the problem of recognizing dialectal or foreign accented data. The proposed methods vary from simple collecting data in the target accent and training new acoustic models, to various ways of adapting models trained on unaccented speech to the new accent. Wang *et al.* [1] investigated German-accented English speakers while Tomokiyo and Waibel in [3] examined Japanese-accented English speakers for two different tasks. In both cases, it was shown that training on non-native speech data achieves the biggest gains in performance on accented data. The simplest use of adaptation was based on direct use of maximum likelihood linear regression (MLLR) to adapt individually to each test speaker or to a class of accented speakers. In [4] standard MLLR was also used to adapt a Mandarin system trained on speakers from the Beijing area, to recognize Shanghainese-accented Mandarin speakers.

In the above work, a general method to deal with accent is to adapt prior models to the new accent. When multiple accents are present ([2, 5]), cross accent experiments show that performance of accent-independent systems is significantly worse than that of accent-dependent ones, thus the goal is to build multiple models of smaller accent variances, and then use a model selector for the adaptation. Prior accent identification research also mostly focuses on the foreign (non-native) accent problem. Teixeira *et al.* [6] proposed a Hidden Markov Model (HMM) based system to identify English with 6 foreign accents. A context independent HMM was used since the corpus consisted mostly of isolated words, which is not usually the case in tasks of interest. Hansen and Arslan [7] also built HMM to classify foreign accent of American English. They analyzed some prosodic features impact on classification performance and concluded that carefully selected prosodic features would improve the classification accuracy. Instead of phoneme-based HMM, Fung and Liu [8] used phoneme-class HMMs to differentiate Cantonese English from native English. Berkling *et al.* [9] added English syllable structure knowledge to help recognize 3 accented speaker groups of Australian English. Huang *et al.* [5], Chen *et al.* [10] and Zheng *et al.* [2] addressed the problem of identifying native multi-accented Mandarin using Gaussian Mixture Models (GMMs) as accent classifiers.

In this paper, we examine the effect of accent variation in English broadcast news data collected from various geographical regions, where English is spoken as an official language. We first demonstrate the effect of unseen accents on previously trained automatic speech recognition system. We then examine two ways for compensating for accent variation. The first is to train an accent-independent model on a large corpus collected from all accents. This solution drastically improves the performance of the system across all data, but still results in high Word Error Rate (WER) for some accent subsets. To target further performance improvements for individual accent subsets we explore the use of accent-dependent models. Similar to [6, 10], in order to do accent classification, we train two GMMs for each accent: one for male, the other for female, as gender is an important speaker variability factor. Given the test utterances, the speakers' gender and accent can be identified sequentially.

This paper is organized as follows. In Section 2, we describe the multi-accent corpus used for this task. The description of the baseline system and training/testing procedures used for our experimental setup is presented in Section 3, where the baseline accent-unaware system is compared to an accent-independent system. In Section 4 we investigate the performance of our GMM accent classifier, and report accent-dependent recognition results. Section 5 concludes with sum-

[*] This work was performed while visiting LIMSI in 2010 as a guest researcher under the Digiteo program.

[†] This work is partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation and by the ANR EdyLex project.

|  | US | AU | GB | NA | ME | IN | All-nonUS |
|---|---|---|---|---|---|---|---|
| Training shows | 667 | 461 | 225 | 72 | 34 | 26 | 818 |
| Training hours | 316 | 33 | 55.4 | 27.7 | 8.2 | 9.4 | 133.7 |
| Training words | 3.7M | 383K | 660K | 320K | 93K | 115K | 1.57M |
| Test shows | 10 | 4 | 3 | 1 | 1 | 1 | 10 |
| Test minutes | 172 | 12 | 48 | 15 | 13 | 15 | 103 |
| Test words | 29433 | 2173 | 5971 | 2515 | 2005 | 2532 | 15196 |
| Test speakers | 202 | 19 | 45 | 15 | 15 | 15 | 109 |

Table 1: *Multi-Accented English Broadcast News data corpus used in this work. (US): United States, (AU): Australia, (GB): Great Britain, (NA): North Africa, (ME): Middle East, (IN): India.*

mary of our work and discussions on possible future extensions.

## 2. Data

We used a corpus of broadcast news shows from 6 different English speaking regions. The data distribution for the training and test accent subsets is shown in Table 1, which includes the number of shows, hours, words and speakers (only for test data) for each subset. The size of the testsets was selected based on the amount of training data available for each accent. In the training data, there was about 70% male speech in all accent subsets, except GB and NA, where the two genders were equally distributed. For all test sets, there was between 40-70% male speech, except in NA where it was more than 90% male.

The US part of the training data is broadcast data available by the Linguistic Data Consortium (LDC) (Hub4 and TDT corpora). The rest of the data (including the US portion of the test set) was collected in several projects and transcribed by partners in them. The audio comes from a variety of news sources (ABC, Skynews, BBC F24 Euronews, ITV1 etc.) and was mostly collected via satellite with some downloaded from the web. In particular, all of the Indian data comes from the web.

## 3. Baseline system

The speech transcription system uses the same basic modeling and decoding strategy as in the LIMSI English broadcast news system [11].

The acoustic features are derived from a PLP-like [12] acoustic parameterization, which has been used in the LIMSI systems since 1996. The speech features consist of 42-dimensional feature vector. The first 39 features consist of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives, derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. These cepstral coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. A 3-dimensional pitch feature vector (the pitch and delta and delta-delta pitch) is appended to the above cepstral parameters. The pitch contour is derived using the ESPS with the CU interpolation (-D option) and the running average.

Each phone model is a tied-state left-to-right, 3-state continuous density HMM with 32 gaussian components. The silence model uses 2048 gaussians. The triphone contexts to be modeled are selected based on their frequencies in the training data, with a backoff by merging contexts for infrequent triphones. We use a total of about 360k gaussians (17K phone contexts with 11600 tied states). The acoustic models are gender-dependent, speaker-adapted and trained with the Maximum Mutual Information Estimation (MMIE) criteria.

The language model (LM) training corpus is comprised of 1.2 billion words of texts from various LDC corpora (English Gigaword, BN transcriptions, commercial transcripts), news articles downloaded from the web, and internal transcriptions. The LMs are interpolated backoff n-gram models estimated on subsets of the available training texts. A 65k recognition word list, which includes several thousand compound words and acronyms, was selected by interpolation of unigram language models, each trained on a subset of the language modeling training texts so as to minimize the out-of-vocabulary (OOV) rate on a set of development data.

The transcription system has two main components, an audio partitioner and a word recognizer. Data partitioning is based on an audio stream mixture model [13], and serves to divide the continuous stream of acoustic data into homogeneous segments, associating cluster, gender and labels with each non-overlapping segment. For each speech segment, the word recognizer determines the sequence of words, associating start and end times and an optional confidence measure with each word.

Word recognition in this work was performed in a single real-time decoding pass, generating a word lattice with cross-word, position-dependent, gender-dependent acoustic models, followed by consensus decoding [14] with 4-gram and pronunciation probabilities. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [15] techniques prior to decoding.

The first line in Table 2 shows the Word Error Rate (WER) performance of a previously available broadcast news system, trained on US broadcast news data as described in [16], on the different datasets used in this work. The acoustic training data used for that system included 180 hours of LDC Broadcast News data and 450 hours of LDC TDT4 data with light supervised transcriptions. We see that for all accents the results are much worse than for US, and in some cases the WER is more than 3 times higher.

In the second line of the same table we see the results with an accent independent acoustic model. In order to maintain a more balanced ratio between US and other accents, only a portion of the original US data was used along with the multi-accented training corpus. The amount of hours from each accent used for training is shown Table 1. The acoustic model size (total number of gaussians) remained the same. We see a reduction of WER to about half, for the datasets from AU, GB and NA and a significant reduction for ME and IN. We observe that even on the US-portion of the testset the WER is slightly improved, due to the increased variability in the training data, even though the total amount of training data hours did not increase.

| | US | AU | GB | NA | ME | IN | Sum | Ave |
|---|---|---|---|---|---|---|---|---|
| US-only baseline | 15.32 | 21.86 | 24.63 | 33.00 | 43.77 | 55.45 | 21.44 | 41.43 |
| Accent independent | 14.34 | 11.92 | 12.84 | 15.90 | 26.47 | 39.28 | 16.07 | 20.12 |
| US-adapt | **13.95** | 15.75 | 16.17 | 22.88 | 31.41 | 46.35 | 17.46 | 24.42 |
| AU-adapt | 17.20 | **11.91** | 13.80 | 19.79 | 32.22 | 45.86 | 18.93 | 23.46 |
| GB-adapt | 17.47 | 12.89 | **11.98** | 16.65 | 29.58 | 42.11 | 18.41 | 21.78 |
| NA-adapt | 16.49 | 13.21 | 12.60 | **16.46** | 27.33 | 40.52 | 17.66 | 21.10 |
| ME-adapt | 16.05 | 12.43 | 13.66 | 19.34 | **25.19** | 39.89 | 17.50 | 21.09 |
| IN-adapt | 16.13 | 14.27 | 14.01 | 19.85 | 27.97 | **34.28** | 17.53 | 21.08 |
| Accent-aware | 13.95 | 11.91 | 11.98 | 16.46 | 25.19 | 34.28 | 15.39 | 18.96 |

Table 2: *WER results using different acoustic models for recognition. Results are reported on each of the 6 regional subsets: United States (US), Australia (AU), Great Britain (GB), North Africa (NA), Middle East (ME), India (IN). The Sum result corresponds to the overall WER on the whole testset, while the Ave result is the average WER when each subset is weighted equally. The first part compares the baseline (US-only trained) model with the newly trained accent independent model. For the second part accent-dependent (MAP adapted) acoustic models are used. The final Accent-aware result is obtained by using for each show the accent-adapted model corresponding to the known show's region of origin.*

# 4. Accent-dependent recognition

## 4.1. Accent Adaptation

As seen in Table 2, the accent independent system still has quite a high WER for two of the accents, ME and IN, which are least represented in the training data. In order to obtain models better matched for each accent we used MAP adaptation to adapt the accent independent acoustic model to the data available for each accent. We used the algorithm in [17] to adapt the gender-independent accent-independent maximum-likelihood trained models to the gender-specific and accent-specific training subsets, in order to obtain gender and accent dependent HMMs. The weight for the adaptation data was set to 10 - we tried other weights but the results didn't change much. We also experimented with doing the adaptation in two steps, first adapt to gender and then to accent, and also with doing gender-independent adaptation, but the joint accent+gender adaptation strategy gave the best overall results. Following MAP adaptation, we performed one iteration of MMIE training (as was the case with the accent-independent model), using the original lattices from all accents and genders. We experimented with MMIE training using accent-dependent data for each accent-adapted model, but the results were worse. Overall the improvements we got from MMIE after MAP were small (varying between 1-6% relative for different subsets). We did not try using MAP adaptation on top of MMIE trained models, using MMIE-MAP as in [18], but this approach could also be explored in the future.

The results of the accent-dependent models for each test subset are shown in the second part of Table 2. We notice that, compared to the accent-independent result, there is a significant improvement in performance for ME and IN data when accent-matched models are used, a small improvement for US and GB, no change for AU, while the performance actually gets worse for the NA data. The accent aware result in the final line of the table is obtained by selecting for each show the accent-dependent model corresponding to the known region of origin for the show.

## 4.2. Accent Identification

A GMM based classifier was used to perform accent identification. We first trained a global GMM with 2048 gaussian mixtures from all training data. Then this GMM was MAP-adapted to the gender specific data for each accent, to get gender+accent dependent GMM models. Accent identification was performed after the speech and gender partitioning step of the speech recognition system. The likelihood for each speech segment was computed using all accent GMMS for the identified gender. The decision for the accent was made either across each speaker-cluster or across each show (whole news file). We found that making the decision based on the average score for each accent-GMM across all speech segments of the speaker-cluster or show, was giving slightly better results than using the overall likelihood score (summing over all speech segments). This strategy gave higher classification accuracy both on training and test data. The testset results for confusion segments and precision/recall presented in Table 3 are given for decisions based on average per segment scores, with the decision made for each speaker cluster. When the decision was made for each show, we had 100% classification accuracy both on the test and training data.

## 4.3. Word Recognition with Automatic Accent ID

In Table 4 we present the WER results when recognizing each speech segment with the identified accent-dependent model. Since we achieve 100% show classification accuracy the results with show-accent-ID are the same as with the accent-aware result in Table 2. We observe that for US and GB, the speaker level accent id leads to worse results, while for the other accents it gives about the same performance.

# 5. Discussion

In this work we investigated the effect of accent variation in English broadcast news data collected from various geographical regions. We found a drastic performance degradation when a system trained on a single accent (US-only) was used to recognize data from other regions. An accent-independent acoustic model, trained on a mixture of data from all accents, achieves a good performance overall on all data. Using GMM-based show-level accent identification we were able to achieve further improvements.

Even though show-level accent-ID achieves good results on average across shows, there is indication that results can be further improved. For example, on the NA dataset the result of the accent-aware system is worse than that with the accent

|  | US | AU | GB | NA | ME | IN | Precision | Recall |
|---|---|---|---|---|---|---|---|---|
| US | 7361 | 425 | 24 | 586 | 562 | 981 | 0.93 | 0.72 |
| AU | 87 | 564 | 55 | 26 | 0 | 0 | 0.42 | 0.77 |
| GB | 472 | 170 | 1376 | 497 | 34 | 129 | 0.80 | 0.52 |
| NA | 56 | 0 | 0 | 805 | 0 | 0 | 0.42 | 0.93 |
| ME | 24 | 0 | 138 | 31 | 487 | 95 | 0.45 | 0.63 |
| IN | 2 | 0 | 0 | 0 | 0 | 859 | 0.41 | 0.99 |

Table 3: *Classification confusions (in seconds) and % precision and recall for speaker-cluster accent classification. The true labels for each speaker is unknown, so we use as target the accent of the show origin.*

|  | US | AU | GB | NA | ME | IN | Sum | Ave |
|---|---|---|---|---|---|---|---|---|
| speaker-accent-ID | 14.71 | 11.23 | 13.10 | 16.46 | 25.34 | 34.28 | 16.01 | 19.18 |
| Show-accent-ID | 13.95 | 11.91 | 11.98 | 16.46 | 25.19 | 34.28 | 15.39 | 18.96 |

Table 4: *WER results for the different subsets of the testset, using the accent-dependent (MAP adapted) acoustic models after speaker-level and show-lever accent ID*

.

independent model, which may indicate mixed accent data in that dataset. The use of finer granularity for accent-ID (e.g. speaker-level) has the potential of capturing speaker accent variation within a show. It is possible that a more accurate accent classifier has to be used in order to see improvements with speaker level accent-ID. Future work involves exploring the use of phonotactic-based classifiers that are commonly used for language ID tasks. Accent classifiers can be used to automatically group the training data as well, and models can be adapted on subsets of training data that can span across geographic regions. Furthermore, automatic clustering of training data across accent regions can exploit accent similarities to improve the results of one accent using data from another.

Finally, this work explores only one approach, acoustic model adaptation, to compensate for accent variation. Future work with this data will examine also the use of pronunciation and LM adaptation, in combination with different approaches for acoustic model adaptation, for a more robust system performance across different English Broadcast News sources.

## 6. References

[1] Z. Wang, T. Schultz, and A. Waibel, “Comparison of acoustic model adaptation techniques on non-native speech,” in *Proc. ICASSP*. IEEE, 2003, pp. 540–543.

[2] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S. youn Yoon, “Accent detection and speech recognition for Shanghai-accented Mandarin,” in *Proc. Interspeech*, 2005.

[3] L. M. Tomokiyo and A. Waibel, “Adaptation methods for non-native speech,” in *Proc. of Multilinguality in Spoken Language Processing*, 2001.

[4] C. Huang, E. Chang, J. Zhou, and K.-F. Lee, “Accent modeling based on pronuntiaion dictionary adaptation for large vocabulary Mandarin speech recognition,” in *Proc. ICSLP*, vol. 2, 2000, pp. 818–821.

[5] C. Huang, T. Chen, and E. Chang, “Accent issues in large vocabulary continuous speech recognition,” *International Journal of Speech Technology*, vol. 7, pp. 141–153, 2004.

[6] C. Texeira, I. Trancoso, and A. Serralheiro, “Accent identification,” in *Proc. ICSLP*, 1996, pp. 1784–7.

[7] L. M. Arslan and J. H. Hansen, “Frequency characteristics of foreign accented speech,” in *Proc. ICASSP*. IEEE, 1997, pp. 1123–1126.

[8] P. Fung and L. W. Kat, “Fast accent identification and accented speech recognition,” in *Proc. ICASSP*. IEEE, 1999, pp. 221–224.

[9] K. Berkling, M. Zissman, J. Vonwiller, and C. Cleirigh, “Improving accent identification through knowledge of English syllable structure,” in *Proc. of ICSLP*, 1998, pp. 89–92.

[10] T. Chen, C. Huang, E. Chang, and J. Wang, “Automatic accent identification using Gaussian mixture models,” in *Workshop on ASRU*. IEEE, 2001, pp. 343–346.

[11] J. L. Gauvain, L. Lamel, and G. Adda, “The LIMSI broadcast news transcription system,” *Speech Communication*, vol. 37, pp. 89–108, 2002.

[12] H. Hermansky, “Perceptual linear prediction (PLP) analysis for speech,” *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[13] J. L. Gauvain, L. Lamel, and G. Adda, “Partitioning and transcription of broadcast news data,” in *Proc. ICSLP*, 1998, pp. 1335–1338.

[14] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus among words: Lattice-based word error minimization,” in *Proc. Eurospeech*, 1999, pp. 495–498.

[15] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[16] L. Lamel, J. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, “The LIMSI 2006 TC-STAR EPPS transcription systems,” in *Proc. ICASSP*. IEEE, 2007.

[17] J. L. Gauvain and C. hui Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[18] D. Povey, M. G. amd D.Y. Kim, and P. Woodland, “MMI-MAP and MPE-MAP for acoustic model adaptation,” in *Proc. Eurospeech*, 2003, pp. 1981–1984.