# Effect of Foreign Accent on Speech Recognition in the NATO N-4 Corpus

*Aaron D. Lawson[1], David M. Harris[2] and John J. Grieco[3]*

[1]Research Associates for Defense Conversion, Rome, NY USA
[2]ACS Defense, Inc., Rome, NY USA
[3]Air Force Research Laboratory, Rome, NY, USA
Aaron Lawson@rl af mil
David Harris@rl af mil
John Grieco@rl af mil

## Abstract

We present results from a series of 151 speech recognition experiments based on the N4 corpus of accented English speech, using a small vocabulary recognition system. These experiments looked at the impact of foreign accent on speech recognition, both within non-native accented English and across different accents, with particular interest in using context free grammar technology to improve callsign identification. Results show that phonetic models built from foreign accented English are *not* less accurate than native ones at decoding novel data with the same accent. Cross accent recognition experiments show that phonetic models from a given accent group were 1.8 *times* less accurate in recognizing speech from a different accent. In contrast to other attempts to perform accurate recognition across accents, our approach of training very compact, accent-specific models (less than 3 hours of speech) provided very accurate results without the arduous task of adapting a phonetic dictionary to every accent.

## 1. Introduction

The Air Force Research Laboratory (AFRL) Speech and Audio Processing Lab had four main goals in undertaking these experiments:
1) To understand and evaluate the impact of foreign accented English on speech recognition using data that emulates tactical conditions.
2) To verify whether accurate phonetic models could be trained from very small, foreign accented audio data-sets without adjusting components of the phonetic recognition system for the accent in question.
3) To determine the effectiveness of phonetic models trained from one accent at decoding a different accent.
4) To demonstrate the effectiveness of targeting the grammar component of our speech recognizer to identify strings of special importance (in this case naval callsigns) by incorporating a context free grammar representation of all possible callsigns into grammar training.

Studies have certainly shown that native models perform better on native-accented language decode data than they do on foreign accented data. A significant amount of work has been done in adapting native English phonetic models from large vocabulary ASR systems to perform better on accented data, generally through detailed phonetic re-transcription of data, with decent results [2], [3]. However, since this process is very time-consuming, labor-intensive and requires a great deal of training, it is simply not feasible in the world of NATO military audio. Livescu [4] proposes building a phonetic model that combines a proportion of native and non-native accented speech, in addition to directly modeling the nature of foreign accented speech in the training. This is problematic in that the improvement in accuracy obtained through the combined-model approach is only slight, and, while the direct modeling approach does improve accuracy considerably, it has the same feasibility drawbacks of [2] and [3]. Our intent in this study is to propose a fast and flexible approach to improving accuracy on accented data, and to demonstrate its application in identifying key audio elements.

## 2. Description of System

The recognition experiments were run using a continuous speech recognition engine developed for AFRL. This is a single-pass small vocabulary recognizer adapted from the Byblos recognizer. The triphone based phonetic models used for these experiments were trained entirely from the limited data available in the NATO N4 corpus. These models used the CMU phonetic dictionary, without any accent specific modifications.

The stochastic grammar model normally used in this ASR system was augmented with a context free grammar (CFG) capability to model callsigns. For each target accent, a model of the callsign syntax was created using a CFG that described the possible callsigns used by that Naval group, without any probability information on the specific alpha-numerics in the corpus itself. For example, the callsign "BRAVO SIX CHARLIE" was modeled as "<ALPHA> <DIGIT> <ALPHA>", uniformly introducing an element of

confusability into the language models and allowing them to generalize to all possible callsigns, even ones they have never encountered.

For all of the experiments, a strict separation between training and testing data was maintained. Due to the limited size of the corpus, this necessitated that the experiments be conducted in a round-robin fashion. The speakers from each of the accents were divided into a number of subsets, with each speaker subset in turn used as the test set. The phonetic and language models for each target speaker subset were trained without reference to the audio or transcripts from any speaker in the subset.

## 3. Data

The NATO Research Technology Organization (RTO) Information Systems Technology Panel (IST) Task Group 001 (TG001) 'Speech and Language Technology" created a corpus for the study of non-native accents. The database is the NATO Native and Non-Native N4 corpus. The database is completely described with contact information on how to obtain it in [1]. The database was collected in four countries The United Kingdom, Germany, Canada, and The Netherlands. The recordings are primarily of NATO naval communications training sessions in English. In addition to the naval activity each speaker read the text of the 'North Wind and the Sun" in both English and native language in the case of Germany (German), The Netherlands (Dutch), French-Canadian (French).

*Table 1:* Duration of the database in hours

|              | CA  | GE  | NL  | UK  | ALL |
|--------------|-----|-----|-----|-----|-----|
| Signal       | 5   | 3.5 | 5   | 6.5 | 20  |
| Silence      | 3   | 0.5 | 2   | 4.5 | 10  |
| Speech       | 2   | 3   | 3   | 2   | 10  |
| Naval        | 2   | 2.5 | 2   | 1.5 | 8   |
| Read Passage | 0.3 | 0.7 | 0.7 | 0.3 | 2   |
| Non-Native   | 0.3 | 0.4 | 0.3 | 0   | 1   |
| Native       | 0.3 | 0.3 | 0.4 | 0.3 | 1   |

The audio was recorded with DAT recorders and converted to 16KHz 16-bit linear PCM. The audio was annotated with the Transcriber tool where both full transcription and speaker markings have been completed. The duration of the raw collection ranged per country from 3 hours to 6 hours for a total of 20 hours. The duration of actual speech ranged from 2 to 3 hours per country for a total of around 10 hours. The tactical Naval activity accounted for a total of 8 hours and the read speech component accounted for 2 hours. Table 1, above, summarizes these durations.

*Table 2:* Speaker Information

|            | CA    | GE    | NL    | UK    | ALL   |
|------------|-------|-------|-------|-------|-------|
| # Speakers | 22    | 51    | 31    | 11    | 115   |
| # Women    | 5     | 0     | 9     | 7     | 19    |
| Age        | 22-35 | 17-23 | 17-61 | 19-62 | 17-64 |
| Age Mean   | 28    | 20    | 21    | 28    | 23    |

There is a large degree of accent variability within each country. The total number of speakers also varies from 11 to 51 for a total of 115. The average age is 23 years old and the percentage of women in the dataset is 20%. Table 2 summarizes the corpus with regard to speakers. As this data is well suited for automatic speech recognition and incorporates non-native language issues it is an excellent choice for the focus of this paper.

## 4. Procedure

Our approach was to first benchmark recognition accuracy within each accent group (intra-accent), and then use the most accurate phonetic model from each accent as a basis for performing experiments across accents (inter-accent). For the intra-accent set of experiments each accent group was divided up into training and testing subsets. Ten balanced subsets were created for each accent (eight for UK) with approximately 90% of the data used to train models and 10% used to decode for each subset. The subsets were devised in such a way that speakers in the training set were never present in the decode data for a given set. For example, in a given group one subset might include speakers 1 through 28 for training and speakers 29 to 35 for testing. Phonetic Models (PM) and Grammar Models (GM) were build with each of the training sets, and each model was tested on the speakers that were excluded from training. Thus, 38 models were built and tested in this stage. These thirty-eight models were evaluated for word accuracy, callsign accuracy and keyword (callsign alpha-numeric component) accuracy.

The second set of experiments used the best performing PM from each accent group to decode all of the speaker subsets from the three other accent groups. For these experiments grammar models were build from the decode language transcripts, minus the utterances of the speakers being tested. In total 114 experiments were run and evaluated for word accuracy, callsign accuracy and keyword accuracy.

## 5. Results/Analysis

Please note that each row in tables 3 through 6 represents the average of eight to ten experiments, depending on the number of subsets decoded for each accent group. Table 3 gives the NIST evaluation values for speech recognition for phonetic models decoding data within their own accent, Table 4 provides the same information for each cross-accent pair. Tables 5 and 6

give word accuracy measures for what we refer to here as 'keywords': the alpha-numeric components that make up callsigns, e.g. alfa, bravo, four, etc. Additionally, these two tables show accuracy measures for whole callsigns, with a correct decode requiring that the entire callsign string in the audio, 'alfa zero zulu', for example, be found. Any deviation from the truth (i.e. deletions or insertions) resulted in counting the entire callsign as an error.

*Table 3*: Average intra-accent word accuracy results

|          | COR   | SUB   | DEL   | INS  |
|----------|-------|-------|-------|------|
| German   | 73.74 | 13.48 | 12.75 | 2.27 |
| British  | 69.45 | 16.13 | 14.43 | 3.05 |
| Dutch    | 81.19 | 11.77 | 7.05  | 2.81 |
| Canada   | 74.34 | 16.85 | 8.8   | 5.08 |
| **Average** | **74.68** | **14.92** | **10.09** | **3.65** |

In table 3 we find that the accuracy of foreign-accented English phonetic models is *not* lower than those built from native data, despite the fact that the dictionaries used in phonetic representations were designed for native pronunciations. In fact, the native set (British) was lower than the others, though this may reflect the slightly smaller amount of audio and less speakers available in UK English.

*Table 4*: Average cross-accent word accuracy results

|                      | COR   | SUB   | DEL   | INS   |
|----------------------|-------|-------|-------|-------|
| Canada PM on British | 43.68 | 38.38 | 17.98 | 8.08  |
| Canada PM on Dutch   | 48.37 | 44.14 | 7.44  | 15.57 |
| Canada PM on German  | 47.37 | 41.53 | 11.06 | 9.61  |
| Dutch PM on Canada   | 43.9  | 39.51 | 16.62 | 5.01  |
| Dutch PM on British  | 34.49 | 40.95 | 24.53 | 3.51  |
| Dutch PM on German   | 50.75 | 33.43 | 15.83 | 4.78  |
| British PM on Dutch  | 36.18 | 55.36 | 8.48  | 15.05 |
| British PM on German | 30.8  | 51.29 | 17.93 | 6.48  |
| British PM on Canada | 41.06 | 43.77 | 15.18 | 5.78  |
| German PM on Dutch   | 49.99 | 39.23 | 10.78 | 7.91  |
| German PM on British | 28.54 | 42.4  | 29.06 | 3.46  |
| German PM on Canada  | 40.11 | 42.47 | 17.45 | 5.17  |
| **Average**          | **41.27** | **40.9** | **15.52** | **8.36** |

These results support the viability of our approach of retraining a small phonetic model to capture the specifics of an accent group, without the time-consuming process of recreating a precise dictionary that reflects the specifics of a particular accent.

The results in table 4 clearly show that word accuracy is always reduced when a model built from audio with one accent is used to decode a different accent, by an avergage of 33.4 percentage points (74.7% to 41.3% or 1.8 times more accurate for same-accent decode). To insure that this discrepency was not the result of interference from channel and other acoustic conditions we ran a set of tests using the TED (Translanguage English Database) corpus [5], a dataset of foreign-accented English collected at the Eurospeech conference and available from the Linguistic Data Consortium. These tests corroborated the N4 results almost exactly, with same accent decodes being 1.7 times more accurate than cross accent decodes.

In general, models built from languages that are closely related phonologically (German and Dutch) were more accurate at decoding one another. As one can see, cross-decodings between Dutch and German averaged 50.4% word accuracy, while all other models averaged 39.5%, a difference of 11 percentage points. The native model did not perform better at cross-accent speech recognition than did the non-natives, and there is no reason to suppose that it would form a better base for improving recognition on accented speech than non-native models.

*Table 5*: Average keyword and callsign accuracy in intra-accent experiments

|          | Keyword Accuracy | Callsign Accuracy |
|----------|------------------|-------------------|
| Canada   | 90.69            | 77.85             |
| British  | 80.66            | 63.73             |
| Dutch    | 94.73            | 87.43             |
| German   | 89.73            | 81.27             |
| Average  | **88.95**        | **77.57**         |

As would be expected, gender of speakers had a major impact on recognition performance across sets: a phonetic model built with a majority of female speakers (British subset) performed most poorly (30.8% word accuracy) on German, which is the only subset with no females; likewise the German model performed most poorly on the British model (28.5% word accuracy). On average these cross-decodes were 14 percentage points lower than the rest of the inter-accent decodes (29.7% to 43.6%).

As table 5 reveals, the system has a greater degree of accuracy on keyword (callsign components) than it does on average decode (74.7% to 89.0%) for same-accent decodes. This is probably due in part to the use of a specific context free grammar modeling of callsigns intended to allow for a greater generalizability of the grammar model in this critical area. Likewise, whole callsign accuracy is very high, exceeding the accuracy of the average word. This is quite surprising considering the fact that the grammar has no bias towards certain callsign combinations, which probably would have resulted from a purely stochastic grammar training.

*Table 6*: Average keyword and callsign accuracy in inter-accent experiments

|  | Keyword Acc. | Callsign Acc. |
|---|---|---|
| Dutch PM on British | 38.9 | 10.74 |
| Dutch PM on Canada | 51.87 | 21.74 |
| Dutch PM on German | 61.2 | 30.45 |
| Canada PM on Dutch | 55.02 | 17.01 |
| Canada PM on British | 52.19 | 15.1 |
| British PM on Canada | 58.62 | 23.6 |
| British PM on Dutch | 40.73 | 5.44 |
| Canada PM on German | 55.94 | 23.75 |
| British PM on German | 36.09 | 9.62 |
| German PM on Canada | 51.29 | 18.07 |
| German PM on Dutch | 54.32 | 14.75 |
| German PM on British | 35.03 | 5.22 |
| **Average** | **49.27** | **16.29** |

Cross-accent decodes fare very poorly, with same-accent models being almost twice as accurate (89.0% to 49.3%), as can be seen in table 6. This has an enormous impact on our goal of multiword string identification, same-accent models being 4.7 times more accurate at whole callsign identification (77.6% to 16.3%).

## 6. Conclusions

The striking loss in accuracy across accents, especially in callsign identification, and the high accuracy within an accent, demonstrates the usefulness of precise phonetic models trained on specific foreign accents. For many large vocabulary ASR systems, training a phonetic model for a specific accent is untenable, due to the large amount of transcribed audio required and the training time involved. With a small vocabulary system, such as the AFRL customized version of the Byblos recognizer, a phonetic model can be built in less than two hours, using only a few hours of data (average 2.4 in these experiments), and be highly accurate when used on accented data (average 74.7% word accuracy). Unlike other attempts to improve speech recognition on accented English e.g. [2][3][4], which involved laborious manual processes to reconstruct dictionaries reflecting particularities of an accent, this was accomplished by simply retraining the recognizer on accented data. A next step will be to incorporate dialect/accent detection to automatically choose the best phonetic model to use in decoding audio.

## 7. References

[1] Benarousse, L., Geoffrois, E., Grieco, J., Series, R., Steeneken, H., Strumpf, H., Swail, C., Thiel, D., "The NATO Native and Non-Native (N4) Speech Corpus", *Proc. Eurospeech*, 2001.

[2] Humphries, J.J., Woodland, P.C., and Pearce, D., "Using Accent-specific Pronunciation Modelling for Robust Speech Recognition", In *Proc. ICSLP*, 1996.

[3] Humphries, J.J., Woodland, P.C., "Using Accent-specific Pronunciation Modelling for Improved Large-vocabulary continuous speech recognition", In *Proc. EuroSpeech*, 1997.

[4] Livescu, K., *Analysis and modeling of non-native speech for automatic speech recognition*. S.M. thesis, MIT, Cambridge, MA, 1999.

[5] Lamel, L., Schiel, F., Fourcin, A., Mariani, J. and Tillmann, H., "The translanguage English database (TED)". In *Proc.. ICSLP*, 1795-1798, 1994.