Patrol Team Language Identification System for DARPA RATS P1 Evaluation

Pavel Matějka[1], *Oldřich Plchot*[1], *Mehdi Soufifar*[1], *Ondřej Glembek*[1], *Luis Fernando D'Haro*[1]
Karel Veselý[1], *František Grézl*[1], *Jeff Ma*[2], *Spyros Matsoukas*[2], *and Najim Dehak*[3]

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic
[2]Raytheon BBN Technologies, Cambridge, MA, USA
[3]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
matejkap@fit.vutbr.cz, smatsouk@bbn.com, najim@csail.mit.edu

Abstract

This paper describes the language identification (LID) system developed by the Patrol team for the first phase of the DARPA RATS (Robust Automatic Transcription of Speech) program, which seeks to advance state of the art detection capabilities on audio from highly degraded communication channels. We show that techniques originally developed for LID on telephone speech (e.g., for the NIST language recognition evaluations) remain effective on the noisy RATS data, provided that careful consideration is applied when designing the training and development sets. In addition, we show significant improvements from the use of Wiener filtering, neural network based and language dependent i-vector modeling, and fusion.
Index Terms: language identification, noisy speech.

1. Introduction

The goal of the RATS program is to create technology capable of accurately determining speech activity regions, detecting key words, identifying language and speakers in highly degraded, weak and/or noisy communication channels. RATS test and training data are collected under both controlled and uncontrolled field conditions.

The goal of this paper is to describe our Language identification system submitted for the first phase of the Evaluation organized within this project. The primary submission included four systems — three acoustic and one phonotactic:

- JFA (acoustic)
- i-vector-BUT (acoustic i-vector followed by NN)
- i-vector-BBN (acoustic i-vector followed by NN)
- PHN-CZ (phonotactic i-vector extractor)

Our goal in the first phase of the project was to port existing technologies developed for clean data on the data corrupted by non-aditive noise and adapt the system to best fit the data. Our systems make extensive use of subspace projections, mainly in the form of i-vectors [1].

The paper is organized as follows: Section 2 describes the training, development, and evaluation data sets. Section 3 explains the acoustic and phonotactic front-end systems. Section 4 covers the fusion and calibration. Section 5 summarizes the results of individual systems as well as fusion, and section 6 concludes the paper.

2. Data

The Linguistic Data Consortium (LDC) provided the training and test data for the RATS participants. The audio recordings annotated were selected from existing and newly collected data sources as follows:

- Fisher Levantine conversational telephone speech (CTS)
- Callfriend Farsi CTS
- NIST LRE Data - Dari, Farsi, Pashto, Urdu and non target languages
- RATS Farsi, Urdu, Pashto, Levantine CTS

All recordings were retransmitted through 8 different communication channels, labeled by the letters A through H. A "push-to-talk" (PPT) transmission protocol was used in all channels except G. PPT states produce some regions where two or more non-transmission (NT) segments may occur. In addition to the speech (S) and non-speech (NS) regions, these NT regions are supposed to be marked in the annotations.

There are five target languages: Dari, Arabic Levantine, Urdu, Pashto, Farsi, and 10 nontarget languages in four conditions with durations 120s, 30s, 10s and 3s. Only recording from the 120s condition were released for training and development. We therefore had to construct our own development samples for the shorter durations from the 120s audio files, based on BUT's voice activity detection (VAD).

During the development period for the evaluation, LDC delivered three incremental data releases for training and test. Only the first two releases were used for

This work was partly supported by the DARPA RATS Program and by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.Approved for Public Release, Distribution Unlimited.

the main training and development sets. The last one was made available just prior to the evaluation and was used only as training data in one of our final 4 systems. A small official development set was also provided for a "dry run" evaluation, consisting of 747 files, and covering all durations.

We made large effort to partition the data from the first two releases into balanced training and development sets, which are described below: Our *Main development set*:

- Contains all dry run set files
- 2432 files for non target languages, 1000 source files for all target languages except for Dari
- Final set contains 600-900 files for each channel/language
- Approximately 7120 files for each duration

The remaining data were used as our *Main training set*:

- Files where VAD detects more than 60s of speech
- Set contains 30774 files, but unbalanced (668 files for Dari, 12778 for Arabic Levantine)

We have also created *Balanced training set* which is subset of the main one and contain approximately 700 files for every language and also contains files for all durations.

The last training set is *Extended training set*:

- Contains both Main + Balanced sets
- Contains NIST LRE data from 3rd incremental release from LDC
- 30s cuts from 120s condition
- Together 60k files for 120s, 96k files for 30s, 7k for 10s and 3s
- Heavily unbalanced = 4039 files for Dari, 57497 for Arabic Levantine

The evaluation (EVL) data has the following characteristics:

- Number of segments 1914, 1782, 1715, 1340 for 120, 30, 10, 3 sec conditions respectively.
- Balanced with about 300 files per language, only dari has 59 files.
- Approximately 30% of nontarget languages.
- Reference annotation has several problems and needs to be adjudicated.

3. Systems

3.1. Acoustic – JFA

Voice activity detection (BUT VAD) is performed by Neural Network with input consisting of a block of Mel filter outputs with context of 300ms. The NN has 18 outputs: 9 for speech and 9 for nonspeech, each corresponding to one of the channels (source plus 8 retransmitted). HMM with Viterbi decoding is used to smooth out and merge the outputs to speech and nonspeech regions. This NN is trained on RATS data defined for the speech activity detection (SAD) task [2].

Audio files were processed with Wiener filter from Qualcom-ICSI-OGI Aurora front end [3]. The acoustic system used the popular shifted-delta-cepstra (SDC) [4] feature extraction. After discarding silence portions, every 10ms speech-frame is mapped to a 56-dimensional feature vector. The feature vector is the concatenation of an SDC-7-1-3-7 vector and 7 MFCC coefficients (including C0). Cepstral mean and variance normalization, as well as RASTA filtering are applied before SDC. We have used 25 mel-banks and 300-3200Hz bandwidth for MFCC computation. The bandwidth was adjusted to best fit the average spectrograms of the RATS data.

A 2048-component, language-independent, GMM was trained with the EM-algorithm using the balanced training set. We refer to this model as the *Universal Background Model* (UBM).

Joint Factor Analysis (JFA) model [5] without eigenvoices is used to model languages. The $\mathbf{D}$ matrix is initialized by MAP adaptation from UBM with $tau = 10$. We have used 200 dimensional channel matrix $\mathbf{U}$, and linear scoring for deriving 6 final scores (5 for target languages and one for out of set languages). More details about the system can be found in our previous work [5].

3.2. Acoustic – i-vector NN BUT

The feature extraction of this system is similar to that of the JFA system, with two main differences: Wiener filter is not used, and bandwidth for Mel-filter bank computation is 300-3400Hz (as in our NIST baseline).

We used a 2048-component UBM to generate zero and first order statistics which are used for training the i-vector extractor [1]. The output is a 600-dimensional vector for every file.

The final classifier is a three-layer Neural Network [6] where the input is 600 i-vector, 200 hidden layer and 6 outputs (1 nontarget + 5 target languages). Stochastic Gradient training with L2 regularization is used as training procedure. The Neural Network is trained on the Extended training set.

3.3. Acoustic – i-vector NN BBN

BBN estimated i-vectors as described in [1], and adopted neural networks (NN) as LID classifiers. The NNs were trained using the ICSI quicknet NN tools [7] so as to map the i-vectors into language posteriors. Each NN had 3 layers. In order to alleviate the over-fitting problem, 4 NN models were trained with different numbers of hidden nodes (300, 400, 500 and 600), and then were combined by simply taking the arithmetic mean of their output posteriors. The numbers of input and output nodes were equal to the i-vector dimension (400) and the number of languages (6), respectively.

In training, i-vectors were estimated on chunks of speech of approximately 20s in duration, so as to better match the shorter duration conditions in testing. The 20s chunks were generated by grouping adjacent segments of speech (as produced by the BBN VAD) from each audio file. Six sets of i-vectors were extracted. Each set was estimated using a language-dependent background model (LDBM) with 1024 Gaussian components, and for each set the NN models were trained separately. Finally, the NN posteriors from the 6 NN models were combined by taking the geometric mean [1].

3.4. Phonotactic – PHN-CZ

The phone recognizer is based on a hybrid NN/HMM approach, where neural networks are used to estimate posterior probabilities of phonemes from Mel filter bank log energies using the context of 310ms around the current frame. A 4-layer NN is trained on the Czech CTS data where 30% was artificially corrupted with noise at lowest level 10dB.

The recognizer is able to produce phone lattices from which posterior-weighted counts ("soft-counts") were used in the subsequent processing [8]. A low-dimensional multinomial subspace over the trigram counts in the Main training set is trained using the approach described in [9]. We use the multinomial subspace model along with hard pruning of the low-frequency trigrams to overcome the problem of the data sparsity [9]. The i-vectors are the point estimates of the latent variables describing the coordinates of count vectors in the new low-dimensional sub-space model. The output is a 600-dimensional vector for every file.

The final model was discriminatively trained (on the Main training set) via regularized multiclass logistic regression [10]. The input vectors (of dimension 600) were conditioned by within-class covariance normalization (WCCN).

4. Fusion, calibration and decision making

4.1. Pre-calibration

Each of the recognizers was independently pre-calibrated with an affine transform, trained on the main development data-set:

$$\mathbf{r}_t = \mathbf{C}\mathbf{s}_t + \mathbf{d} \tag{1}$$

where $\mathbf{C}$ is a full K-by-K matrix, $\mathbf{d}$ is a K-dimensional vector and K is the number of classes/languages (in our case 6). These parameters were trained by regularized logistic regression. WCCN was applied at this step.

After pre-calibration, zero vectors, $\mathbf{r}_t = \mathbf{0}$, were inserted for those segments for which the basic recognizers failed to produce scores or input vectors.

[1] Results showed that the geometric mean is better than the arithmetic mean in this case

Table 2: *Results for Fusion - target metric for RATS.*

Pmiss at Pfa=10 [%]	120s	30s	10s	3s
Development	0.26	0.77	4.88	27.68
Evaluation	0.72	3.90	9.32	25.81

4.2. Fusion

Let $\mathbf{r}_{ti}$ denote the outputs of the ith pre-calibrated recognizer. These outputs were fused as:

$$\boldsymbol{\ell}_t = \sum_i \alpha_i \mathbf{r}_{ti} + \boldsymbol{\beta} \tag{2}$$

where each α_i is a scalar weight and $\boldsymbol{\beta}$ is a K-dimensional vector. These parameters were again trained by multiclass logistic regression on the development set. Here, neither WCCN, nor regularization, were applied.

The output vector of the fuser can be interpreted as multi-class log-likelihoods for the open-set language identification task. For creating the decisions, these log-likelihoods are converted into log-likelihood ratios with the same priors as defined in the NIST LRE 2009 and the thresholds are selected independently for each duration condition to address the desired operating point defined in the RATS evaluation plan (P_{miss} at $P_{fa} = 10\%$).

5. Results

We used only one development set for calibration and fusion because of the lack of the data for Dari and nontarget languages in the first two releases of data by LDC. But based on BUT and MIT experience from the past NIST LRE we found it is safe to use only one set if it is well designed and big enough. In addition, we did several cross-checks such as: splitting the set to two smaller independent parts, jack-knifing over 5 parts (4 for training, 1 for test). The results were consistent and later were also confirmed on the evaluation data.

Finally, we decided to use a duration-independent calibration and fusion using all the utterances in the development set, except for those from the 3s condition, due to their questionable reliability. This hurt our performance slightly on the 120s condition, but we decided that this was preferrable so as to make the system more robust.

Table 1 presents results of separate sub-systems and final fusion on DEV and EVL data. The EVL data are much harder which might be also caused by annotation errors. Table 2 shows results in terms of target primary metric of the RATS project for Phase 1, which is P_{miss} at $P_{fa} = 10\%$ computed over pooled scores of separate language detectors.

Table 3 shows the effect of using Wiener filter as the preprocessor of audio files. The results are reported on the BUT i-vector system with dimensionality 400 and Logistic regression as classifier.

Table 1: *Results for fusion and individual systems.*

	Development				Evaluation			
Cavg [%]	120s	30s	10s	3s	120s	30s	10s	3s
JFA	1.61	6.14	12.52	23.53	7.05	12.92	17.36	22.68
i-vector BUT	1.60	4.94	10.36	21.73	7.83	9.97	14.52	21.46
i-vector BBN	2.58	5.92	12.06	28.21	9.03	11.96	17.83	27.10
PHN CZ	2.60	8.95	16.84	30.53	9.06	15.56	21.90	29.12
Fusion	0.83	2.92	6.85	18.08	6.56	8.33	11.40	17.45

Table 3: *Results on DEV set for analysis of effect of preprocessing audio data with Wiener Filter.*

Cavg [%]	120s	30s	10s	3s
baseline	2.12	7.06	13.48	24.12
baseline + WF	1.70	6.41	12.96	23.46

Table 4: *Analysis of Neural Network and Logistic Regression as final classifier for BUT i-vectors on DEV set.*

Cavg [%]	120s	30s	10s	3s
LR - Main TRN set	1.86	6.70	13.48	23.60
NN - Main TRN set	1.87	8.09	15.37	24.84
LR - Extended TRN set	2.24	7.22	13.59	24.62
NN - Extended TRN set	1.60	4.94	10.36	21.73

Table 4 shows results with different classifiers: Logistic regression and Neural network on different sizes of training data. The results show that LR is superior when training on our Main set, but if we use Extended set with more data and shorter utterances the NN can benefit from this mainly for short ones, while LR is not able to capture the information. One possible reason is that NN has more parameters to train.

The first line in the Table 5 refers to the baseline of the BBN i-vector system with UBM 1024G, 600 dimensional i-vector and NNs. The following line shows the effect of training i-vector with 20s buffer which improves a lot the results on short durations. The last line presents the use of the LDBMs (language dependent background models) which improve the results by 20-30% relatively [2].

6. Conclusions

We have described the four systems that were part of the Patrol Team Language Identification system for the DARPA RATS project. The main conlusion is that techniques developed on telephone speech mainly for past NIST LRE evaluations remain effective on the RATS radio communication data, provided that the systems are trained on audio from multiple channels. We also found that the cosine distance and Logistic regression do not achieve as good results compared to neural networks,

[2] In this case 400 dimensional i-vectors are used due to the increased complexity of the system.

Table 5: *Cavg[%] mesured on the DEV set, of different BBN systems*

Cavg [%]	chunk size	120s	30s	10s	3s
UBM	whole audio	2.65	11.20	22.77	39.07
UBM	20s	3.21	8.05	16.23	33.10
LDBM	20s	2.58	5.92	12.06	28.21

while on NIST data the results are approximately the same. We are still trying to understand this behavior. Finally, we observed big impact on performance from using short utterances in the modeling.

7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," in *IEEE Transactions on audio, speech and language processing*, vol. 19, no. 4, 2011.

[2] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, K. Vesely, P. Matějka, X. Zhu, and N. Mesgarani, "Developing a speech activity detection system for the darpa rats program," in *submitted to Proc. of Interspeech 2012*, Sep. 2012.

[3] L. Burget, S. Dupont, H. Garudadri, F. Grézl, H. Hermansky, P. Jain, S. Kajarekar, and N. Morgan, "Qualcomm-icsi-ogi features for asr," in *Proc. 7th International Conference on Spoken Language Processing*. International Speech Communication Association, 2002, p. 4.

[4] P. Torres-Carrasquillo, E. Singer, M. Kohler, R. Greene, D. Reynolds, and J. Deller Jr., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Sep. 2002, pp. 89–92.

[5] Z. Jančík *et al.*, "Data selection and calibration issues in automatic language recognition - investigation with but-agnitio nist lre 2009 system," in *Proc. Odyssey 2010*, 2010, pp. 215–221.

[6] K. Veselý, L. Burget, and F. Grézl, "Parallel training of neural networks for speech recognition," in *Prof. Text, Speech and Dialogue 2010*, ser. LNAI 6231, vol. 2010, no. 9. Springer Verlag, 2010, pp. 439–446.

[7] I. N. toolkit. [Online]. Available: http://www.icsi.berkeley.edu/Speech/icsi-speech-tools.html

[8] J. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proceedings of Interspeech 2004*, Sep. 2004, pp. 1283–1286.

[9] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "ivector approach to phonotactic language recognition," in *Proceedings of Interspeech 2011*, vol. 2011, no. 8, 2011, pp. 2913–2916.

[10] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2007.