Improving Language Identification Robustness to Highly Channel-Degraded Speech through Multiple System Fusion

Aaron Lawson, Mitchell McLaren, Yun Lei, Vikramjit Mitra, Nicolas Scheffer, Luciana Ferrer, Martin Graciarena

Speech Technology and Research Laboratory, SRI International, California, USA

{aaron, mitch, yunlei, vmitra, nscheffer, lferrer, martin}@speech.sri.com

Abstract

We describe a language identification system developed for robustess to noise conditions such as those encountered under the DARPA RATS program, which is focused on multi-channel audio collected in high noise conditions. Work presented here includes novel approaches to scoring iVectors, the introduction of several new acoustic and prosodic features for language identification, and discriminative file selection approaches to score calibration. Further, we explore the use of Discrete Cosine Transforms (DCT) as a supplement to traditional context modeling with Shifted Delta Cepstrum (SDC) and fusion of multiple iVector systems based on Gaussian backends, neural networks, and adaptive Gaussian backend modeling.

Index Terms: language identification, speech features, iVector scoring.

1. Introduction

The systems described in this paper were designed to meet the RATS program goals of robustness to high levels of noise across a range of channel conditions and channel distortions. The ultimate goal is a system that is sufficiently robust to enable its use in real-world conditions with all the variability and unpredictability that entails. To this end, the SRI SCENIC system was developed based on a multi-faceted set of technologies, adapting the features, context modeling, scoring, and fusion and calibration. The resulting system uses both noise-robust acoustic features, such as Mean Hilbert Envelope Coefficient (MHEC), Power Normalized Cepstral Coefficients (PNCC) and Medium Duration Modulation Cepstral (MDMC), and Subband Autocorrelation Classification (SACC) prosodic features, along with traditional speech features Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) features. For context modeling the use of discrete cosine transforms to contextualize temporal information augments the usual shifted delta cepstrum. Three scoring approaches are fused in the primary system: Gaussian backend (GB), neural network (NN), and Adaptive Gaussian Backend (AGB). AGB [1] was developed under this effort and is especially effective at scoring very short segments in the 3-second to 10-second range, while NN scoring was optimal for longer durations. The calibration set was determined using a dataset refinement technique based on support vector machines (SVMs) [2].

2. Data

The corpus used for all RATS program experimentation was collected by the Linguistic Data Consortium [3]. The goal of this corpus is to provide a set of data characterized by natural noise induced by variations in transmission conditions across eight different channels (A-H). All the data was retransmitted across all the channels and re-recorded, resulting in more than 100,000 files. The core target languages are Levantine Arabic, Farsi, Dari, Pashto, and Urdu. The evaluation sets also contain ten non-target languages, and the full corpus of training data contains more than thirty languages from which researchers may draw.

2.1. Evaluation sets

Results in this paper will focus on the dev-2 development set established under the RATS program. Dev-2 was chosen over dev-1 since it is substantially larger and represents collections made over a much larger time frame, thereby providing a greater diversity of collection conditions and variation. Dev-2 was broken down into test sets with recordings of four durations of active speech (3, 10, 30 and 120 seconds).

2.2. Training sets

A pool of 88,000 files was selected from the full set by excluding those files either actually contained in the evaluation set or which shared a common original recording with a file in the evaluation set. These 88,000 files will be referred to hereafter as the base training set. Since the distribution of data in the base set is skewed towards certain languages (for example, there are thousands more files from Arabic than there are from Dari and Farsi), a second balanced training set was developed using a relatively well-balanced subset of 21,000 files. This set includes all the available Dari and Farsi data, and 3,200 files from each of the other target languages. The remaining data consists of all the files from the ten non-target evaluation languages with some additional files from closely related languages. Two extended sets were also created from the base set.

The first extended set (ext-1) includes all the base files broken up into 30-second segments to provide 240,000 additional training files plus the base set. The second extended set (ext-2) includes both the 30-second segments and an additional 620,000 10-second segments along with the base. These three sets were used for training different components of the system as presented in Table 1.

3. System Description

The SRI SCENIC team goal for system development was to fuse approaches that contribute to accuracy and noise robust-

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. "A" (Approved for Public Release, Distribution Unlimited)

Table 1: Training set breakdown.

System Component	Corpus	No. Files
Universal Background Model	Balanced	21K
iVector Extractor	Balanced	21K
Gaussian Backend	Balanced	21K
Adaptive Gaussian Backend	Ext-1	328K
Neural Network	Ext-2	948K

ness on the development set without tailoring the system specifically to this dataset. Thus, there was no attempt to build duration-dependent subsystems or systems targeted at specific channels contained in the development set.

The resulting system fuses multiple features, both acoustic (MHEC, MDMC, PNCC, MFCC and PLP) and prosodic (SACC), with two different context modeling approaches: discrete cosine transform (DCT) and Shifted Delta Cepstrum (SDC). All systems are modeled using iVectors, which are scored in three different ways: Neural Network (NN), Gaussian Backend (GB) and a technique developed under this effort, Adaptive Gaussian Backend (AGB). Three different fusion systems were selected with the backend scoring regimes detailed in Table 2.

The primary system was chosen based on accuracy across the four durations, combining NN systems, which perform best on long-duration segments, with AGB, which are especially effective on short durations. The GB scoring performed best with the prosodic features and added substantial improvement in fusion with MDMC.

The secondary system uses only three features and only the AGB scoring technique. The goal of this system was to highlight developments made under the RATS program phase two and demonstrate the power of AGB scoring.

The tertiary system uses the same features as the primary but generates scores with the standard GB approach. This contrastive system aims to demonstrate the benefits of fusing multiple complementary scoring techniques in the primary system.

3.1. Features

The systems described in this paper use a combination of noise-robust features developed under this effort along with conventional speech features such as MFCC and PLP. The goal in developing the set of features was to draw on various approaches to noise robustness from SRIs research partners.

This combination, when properly fused and calibrated, provides substantially greater accuracy than any single feature on its own. The following sections briefly describe the features used in the final three systems.

3.1.1. SACC

Subband Autocorrelation Classification (SACC) [4] is a noise-robust speech tracking technique that uses a multi-layered perceptron (MLP) to estimate the pitch tracks. SACC obtains principal components of the autocorrelations of subband speech signals from an auditory filter bank with 24 channels. Principal Component Analysis (PCA) is applied to each of the subbands to reduce the dimensionality of the autocorrelation measure, and the result is fed as an input to a single hidden-layer MLP with 800 neurons. The output pitch estimates from the MLP are smoothed using Viterbi smoothing and result in a frame rate of 10 ms.

3.1.2. MHEC

The Mean Hilbert Envelope Coefficient (MHEC) feature, introduced in [5], is used primarily for noise-robust speaker identification. In the MHEC front end, the speech signal is pre-emphasized and split into 32 channels uniformly spaced in the ERB scale gamma-tone filterbank with cutoff frequencies of 300 Hz to 3400 Hz. The temporal envelope of each subband component is estimated using the Hilbert transform followed by low-pass filtering with a 20 Hz cutoff frequency. The smoothed Hilbert envelope is then analyzed using a 25 ms Hamming window with a 10 ms frame rate. Log compression is performed on the short-term energy signals, followed by DCT transform to generate seven cepstral features.

3.1.3. MDMC

The Medium Duration Modulation Cepstral (MDMC) feature is obtained using a modified version of the algorithm presented in [6]. In the MDMC front-end, the digital speech signal is pre-emphasized (using a pre-emphasis filter of coefficient 0.97) and then analyzed using a 25.6 ms Hamming window with a 10 ms frame rate. The windowed speech signal is split into 34 band-limited signals using a gamma-tone filter bank, spaced equally from 200 Hz to 3750 Hz in the ERB scale. Amplitude modulation (AM) signals are then estimated from the sub-band signals using the Teager non-linear energy operator. The AM power for each sub-band is estimated over its analysis window, which generates the AM power signal at a 100 Hz sampling rate. The resulting sub-band power signals are then power normalized using 1/15 root. DCT is performed on the root compressed power signal and the first seven coefficients (including the C0) are retained.

3.1.4. PNCC

Power Normalized Cepstral Coefficients (PNCCs) are noise-robust acoustic features based on the work presented in [7]. In PNCC, an acoustic digital signal is pre-emphasized (with a coefficient of 0.97) and then analyzed using a 25.6 ms Hamming window with a 10 ms frame rate. A short-time Fourier analysis is performed over the Hamming windowed data, followed by frequency domain gamma-tone filtering using a 30-channel filterbank with cut-off frequencies 133 Hz and 4000 Hz, where the center frequencies of the gamma-tone bank are spaced equally in the ERB scale. In this implementation of PNCC, small power boosting is supported as explained in [8]. Short-term spectral powers are estimated by integrating the squared gamma-tone responses and the resultant is root compressed using a 1/15th root. DCT is performed on the root compressed power signal, and the first seven coefficients (including the C0) are retained.

3.1.5. Traditional Features

Both MFCC and PLP are used in subsystems. As with the other cepstral features used in this research, the standard c0-c6 coefficients are extracted. MFCC and PLP were constrained to a telephone frequency bandwidth of 200 to 3400Hz.

3.2. Context Modeling

3.2.1. Discrete Cosine Transform Context Modeling

DCT is a means of providing temporal context information to the static cepstrum. Experiments were performed to identify the optimal configuration of DCT for language identification, with multiple different configurations evaluated, covering dif-

Table 2: Configuration of Features and backend scorers used in the primary, secondary and tertiary systems.

Feature	Context Modeling	iVector Dimension	UBM Components	Primary Backends	Secondary Backends	Tertiary Backends
MHEC	SDC	400	1024	AGB	AGB	GB
MDMC	DCT	200	512	NN	-	GB
PNCC	SDC	400	1024	AGB	AGB	GB
SACC	DCT	150	512	GB	-	GB
MFCC	DCT	400	1024	NN	AGB	GB
PLP	SDC	400	1024	GB	-	GB

ferent padding sizes from 6 to 15 and differences in the final DCT vector size. Some of the features used in our experiments (e.g., SACC, MFCC and MDMC) were temporally contextualized with DCT, based on the relative benefit to the fusion system. The final configuration padded 10 frames on either side, and selected every other frame, resulting in 21 frames altogether. Since the temporal contextualization of features results in very high dimensions, DCT is performed on each of the feature dimensions and the top 50% of the coefficients are retained. These are appended to the original feature, giving a total feature vector of 79 dimensions.

3.2.2. Shifted Delta Cepstrum Context Modeling

SDC coefficients are extracted for PLP, PNCC, and MHEC on the delta cepstrum using the standard 7-3-1-7 configuration as in [9] and [10]. The resulting SDC vector is appended to the original cepstrum, resulting in a final vector of 56 dimensions.

3.3. iVector System

The iVector extractor was trained using the balanced dataset of 21,000 files. Training was performed in the standard way, as described in [11]. The optimal iVector dimension for most features was 400 with a UBM component size of 1024, however, for MDMC and SACC, the iVector dimensionality and UBM components were reduced.

3.4. iVector Scoring

Three different approaches to scoring the iVectors were employed in this system: Adaptive Gaussian Backend, Neural Network, and Gaussian Backend.

3.4.1. Adaptive Gaussian Backend

The AGB approach was developed in response to the poor performance obtained using the standard GB in the highly noisy data of the RATS corpora [1]. Due to the very large diversity in conditions that are encountered within a given language class the assumption of a normal distribution on which the GB is founded breaks down, and in many cases the variability between channels and conditions may be greater than the variability between languages. The AGB approach mitigates this problem by adapting the language model mean based on a set of training examples selected at test time using a relevance measure. Initial experiments used Euclidean norm to select the data set for each language class, and maintained the top-N most relevant samples per language. Since the top-N was a parameter that needed to be established empirically rather than dynamically, alternative means of determining the adaptation data were explored. The final approach used a Support Vector Machine (SVM) to select and weight the data for each test case, by using the training data for each class as a background and discriminating the test sample from this background. The resulting support vectors are weighted by their coefficients to produce the adapted, test-specific mean. Readers are directed to [1] for more details on AGB.

3.4.2. Neural Network

Neural network is another successful backend for language identification under the iVector framework, especially on RATS data [12]. A three-layer feed-forward NN is trained on the i-vectors and the corresponding class labels. The mini batch gradient descent algorithm with cross-entropy error backpropagation is used to train the NN [13]. A sigmoid function for nonlinearity is used for both hidden and output layers. The weight is updated with momentum to speed up the training and prevent getting stuck in local minima [13].

3.4.3. Traditional Gaussian Backend

The Gaussian Backend represents the state of the art in iVector scoring for language recognition and was widely used in recent NIST Language Recognition Evaluations [14, 15]. For details on GB scoring with iVectors see [11].

3.5. Fusion and Calibration

The approach of data-driven selection using SVMs [2] was utilized to select the cohort used for calibration. This technique, coined 'dataset refinement', involves two datasets: one is a set of potential candidate samples to be refined and the other a set of target data samples containing properties desired in the refined dataset. Using the potential candidates as the SVM background, an SVM is trained to discriminate these candidates against each sample in the target set. A refined dataset is found by taking the top N samples after ranking by the number of times they were selected as a support vector. For the RATS LID system, we used 12k target samples (the test samples) as the potential candidates and used the 30s+ full training set as target samples. The top 6k candidates were selected as the calibration set and offered approximately 12% relative improvement in performance over the full test set.

4. Results

We report results in terms of Cavg calculated as the average equal error (EER) across the five target languages. The primary fusion system results (Table 3) were best in all durations, and particularly better for longer durations where the two neural network systems contributed the most. The secondary system is considerably simpler than the primary, having only three features and a single scoring method. However, it produces competitive results, especially in the 3-second duration.

The tertiary system uses all six of the features, like the primary, but uses only GB scoring, demonstrating the effectiveness

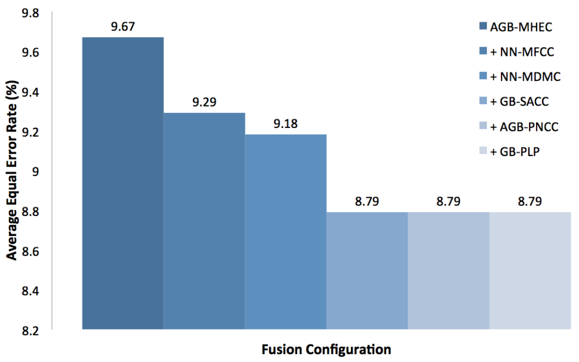

Figure 1: Average EER for 10s tests as next best subsystem of the primary system is added to the fusion pool.

Table 3: Results of fused systems (Cavg).

	Test Duration			
System	**3sec**	**10sec**	**30s**	**120sec**
Primary	15.0%	8.19%	5.67%	2.80%
Secondary	15.1%	8.95%	6.05%	3.71%
Tertiary	18.5%	12.4%	8.43%	5.26%

of fusing scores from multiple backends. Table 4 details the subsystem results for the three different scoring techniques. Of note is the high accuracy of NN scoring for longer durations, having 1.4% absolute lower error for the 120 second MFCC condition over AGB. In contrast, AGB scoring is most effective at shorter durations, with 1.6% absolute lower Cavg at 3 seconds compared with NN on MFCC. While the Gaussian backend systems are never better than either the NN or AGB systems, they do contribute substantially to the primary fusion. To illustrate the extent that these and other subsystems contributed to the primary system, we present in Figure 1 the incremental improvement to the 30s test condition when adding the next best subsystem to the fusion pool. It can be seen that the best performing AGB-MHEC system is complemented by the two NN systems and this fusion is further improved by adding GB-SACC; the worst performing individual system.

5. Conclusion

This paper described the SRI SCENIC language identification system developed for phase two of the DARPA RATS program. Three systems were developed exploiting up to six available features; both traditional and noise-robust features developed under the RATS program. The systems benefitted from both SDC and DCT approaches to contextual modeling as well as three complementary backend scoring regimes: NN, GB, and the novel AGB developed under the RATS program. While NN was shown to perform best on longer test durations, AGB was more suitable for shorter durations. The novel application of SVMs to select the calibration dataset was also found to be effective on the development data. Finally, the contribution of each subsytem of the primary system was highlighted emphasizing the benefit of multiple backends and multiple features in the SRI SCENIC system.

Table 4: Results of individual subsystems (Cavg).

		Test Duration			
Scorer	**Feature**	**3sec**	**10sec**	**30s**	**120sec**
NN	MFCC	21.4%	12.3%	7.8%	**3.7%**
	MDMC	22.5%	12.3%	6.9%	4.1%
AGB	MHEC	**15.1%**	**9.9%**	**6.7%**	4.7%
	PNCC	17.5%	11.5%	7.5%	5.1%
	MFCC	19.8%	11.9%	7.1%	5.1%
GB	MFCC	28.9%	18.3%	12.7%	6.5%
	MDMC	30.1%	19.2%	12.8%	6.8%
	PNCC	26.9%	19.8%	12.9%	7.2%
	MHEC	27.7%	18.7%	12.7%	7.4%
	PLP	30.4%	22.4%	15.7%	9.2%
	SACC	37.4%	33.2%	27.6%	16.7%

6. References

[1] M. McLaren, A. Lawson, Y. Lei, and N. Scheffer, "Adaptive gaussian backend for robust language identification," in *Submitted to Interspeech*, 2013.

[2] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Data-driven background dataset selection for SVM-based speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 18, no. 6, pp. 1496–1506, 2010.

[3] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. The Speaker and Language Recognition Workshop*, 2012.

[4] B. Lee and D. Ellis, "Noise robust pitch tracking by subband autocorrelation classication," in *Proc. Interspeech*, 2012.

[5] J.-W. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin, and J. H. Hansen, "Exploring hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA," in *Proc. NIST Speaker Recognition Evaluation Workshop*, 2011.

[6] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4117–4120.

[7] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4101–4104.

[8] C. Kim, K. Kumar, and R. M. Stern, "Robust speech recognition using a small power boosting algorithm," in *Proc. IEEE Automatic Speech Recognition & Understanding, 2009*, 2009, pp. 243–248.

[9] B. Bielefeld, "Language identification using shifted delta cepstrum," in *Proc. Fourteenth Annual Speech Research Symposium*, 1994.

[10] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. International Conference on Spoken Language Processing*, vol. 2, 2002, pp. 33–36.

[11] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *Proc. Interspeech*, 2011, pp. 861–864.

[12] P. Matejka, O. Plchot, M. Soufifar, O. Glembek, L. F. DHaro, K. Veselỳ, F. Grézl, J. Ma, S. Matsoukas, and N. Dehak, "Patrol team language identification system for DARPA RATS P1 evaluation," in *Proc. Interspeech*, 2012.

[13] C. M. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 4, no. 4.

[14] L. J. Rodríguez-Fuentes, A. Varona, M. Diez, M. Penagarikano, and G. Bordel, "Evaluation of spoken language recognition technology using broadcast speech: Performance and challenges," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012.

[15] M. Penagarikano, A. Varona, M. Diez, L. J. Rodriguez-Fuentes, and G. Bordel, "Study of different backends in a state-of-the-art language recognition system," in *Proc. Interspeech*, 2012.