UTD-CRSS SYSTEMS FOR NIST LANGUAGE RECOGNITION EVALUATION 2011

Gang Liu, Seyed Omid Sadjadi, Taufiq Hasan, Jun-Won Suh, Chi Zhang
Mahnoosh Mehrabani, Hynek Bořil, Abhijeet Sangwan, and John H. L. Hansen

Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering, The University of Texas at Dallas,
Richardson, Texas 75080-3021, U.S.A.
http://crss.utdallas.edu

1. INTRODUCTION

This study summarizes the overall solution and sub-systems developed by the Center for Robust Speech Systems (CRSS) at the University of Texas at Dallas to address the NIST LRE-2011 competition. CRSS-UTD employs five core sub-systems in the proposed language ID solution that include: (1) i-vector, (2) SVM-GSV, (3) PPRLM, (4) Articulatory Feature based, and (5) Prosody based. The first four represent the core solutions, and the fifth represents an investigative effort to incorporate micro-prosodic structure which has previously been explored for dialect ID by CRSS [1], [2]. This paper is organized as follows: first (i) data preparation is discussed for the system development; followed by (ii) System Descriptions; and finally (iii) probe results using LRE-09 data. Score combination of the resulting sub-systems was also considered for overall system development.

2. DATA PREPARATION

Data was prepared separately for conversational telephone speech (CTS) and broadcast narrow-band speech (BNBS). For CTS, the following corpora material was used for building the systems.

- CallFriend,
- Fisher Levantine Arabic,
- LRE'07 Development data,
- OGI-multilingual,
- OGI 22 languages,
- Foreign Accented English,
- LDC2009E42 NIST LRE 2009 CTS Training Data, and
- Evaluation data from LRE96,03,05,07,09 (CTS part).

For BNBS, the following data sources were used:

- Evaluation data from LRE09 (VOA part)
- VOA3 (only audio data used, no label information was used)
- VOA FTP (`ftp://8475.ftp.storage.akadns.net/mp3/voa/`)
- SBS Radio (`www.sbs.com.au`)

For VOA data, music corrupted data was removed using an automatic music detection algorithm based on a spectral flux feature. For all target languages, speaker diarization was performed on the selected segments to ensure that each speaker has an equivalent amount of data existing in the training data.

3. SYSTEM DESCRIPTIONS

3.1. I-vector System

The UTD-CRSS i-vector systems were trained using Mel-frequency cepstral coefficients-shifted delta cepstra (MFCC-SDC) 7-1-3-7 features. Each speech file was segmented into frames of 25 ms duration with 10 ms skip rate. After removing the DC, 24-dimensional mel filterbank log-energies were extracted from hamming windowed frames, and processed with the RASTA filtering. The discrete cosine transform (DCT) was applied to convert the log-spectral coefficients to cepstral coefficients and to decorrelate the various feature dimensions. Only the first 7 coefficients were retained after DCT (including C_0). We employed cepstral mean and variance normalization (CMVN) over a 3-second sliding window to suppress the linear channel effects and to equalize the features. The normalized MFCCs were then converted to MFCC-SDC acoustic features. Since we have observed some room reverberation effects in the training data, as an alternative acoustic feature, we investigated the effectiveness of our recently proposed mean Hilbert envelope coefficients (MHEC) for the LRE task. The MHEC features have been shown to provide robustness in speaker identification under reverberant mismatched conditions [3].

After extracting the acoustic features, training data from all target languages were used to build a 1024-mixture Universal Background Model (UBM).

We followed the i-vector system paradigm for language recognition as presented in [4], [5]. The i-vector system modeling originates from the total variability modeling of speakers proposed in [6]. The main idea of this modeling is to adap the UBM (trained on all available files of all languages) to the set of acoustic feature frames of the test data based on the Eigenvoice adaptation. This adaptation has been previously used to estimate speaker dependent GMM mean supervectors. The main difference in total variability modeling from Eigenvoice modeling is that the training assumes that each utterance comes from a different class. In this view, the modeling assumes that all the variability of a GMM supervector can be summarized by a low rank rectangular matrix T, termed the total variability matrix. Thus, the GMM supervector of a given utterance can be written as:

$$m = m_0 + Tw \tag{1}$$

where, m_0 is the UBM mean supervector and w is the i-vector corresponding to the utterance. In our implementation, we trained the total variability matrix on all the development data (same data used for the UBM) using the EM algorithm for Eigenvoice presented in [7]. We used 5 iterations for estimation. The i-vector extraction pro-

cess is detailed in [6], and we generated 600 dimensional i-vectors in our setup.

3.2. SVM-GSV System

This is a support vector machine (SVM) system using a linear kernel as described in [8]. Gaussian mean supervectors were extracted from MAP adapted GMMs using a small relevance factor. We utilized the SVM classifier from the LIBLINEAR toolkit [9].

3.3. PPRLM System

The implementation of our phone recognition and language modeling (PPRLM) system [10] employs 9 TRAP-based phone recognizers [11] covering the following languages - English, Czech, Hungarian, Russian, German, Hindi, Japanese, Mandarin, Spanish (template languages). The language modeling is conducted using SRILM Toolkit [12]. The architecture of the PPRLM system allows for engagement/disengagement of any of the nine phone recognizers in the decoding phase, which yields 511 (29-1) possible phone recognizer combinations. The optimal subset of the phone recognizers for the language recognition task is searched on the cross-validation set drawn from the small portion of the NIST LRE11 training set (1000 files). For each trial sample, the vector of the PPRLM output likelihoods is normalized with respect to the individual phone recognizer likelihood mean and standard deviation across all target languages as follows:

$$\mu_m = \frac{1}{N}\sum_{\eta=1}^{N} Lik_{m,\eta}, \tag{2}$$

$$\sigma_m = \sqrt{\frac{\sum_{\eta=1}^{N}\left(Lik_{m,\eta} - \mu_m\right)^2}{N}}, \tag{3}$$

$$Lik_{m,n}^{Norm} = (Lik_{m,n} - \mu_m)/\sigma_m, \tag{4}$$

where m is the index of the mth template language (i.e., one of the nine languages modeled by the BUT phone recognizers) and n is the index of the target language (one of the twenty four NIST LID 2011 languages).

3.4. Combined Articulatory and Prosody System

The articulatory system uses a combination of phonological features (PFs) and prosody features. The phonological features are extracted using the process outlined in [1]. Particularly, the following articulatory characteristics were captured: (i) height-of-tongue (height), (ii) frontness-of-tongue (frontness), (iii) lip-rounding (rounding), (iv) nasalization (nasality), (v) excitation (glottal), (vi) place-of-articulation (place), and (vii) manner-of-articulation (degree). Using these articulatory characteristics, Language Features (LFs) are extracted from dynamic PF streams [1]. In the dynamic representation, only the changes in the values of PF-types are of interest. Particularly, both the nature of the change as well as the order in which it occurs are important. In addition to PFs, prosody features were also extracted. Particularly, pitch/energy contour segments were approximated with Legendre polynomial and used as features. N-gram combinations of pitch and energy features were also used for classification. Finally, a maximum entropy classifier (ME) was employed to perform language recognition using the PF and prosody based language features.

4. SCORE FUSION

Experiments were performed on each sub-system using NIST LRE-09 data to determine their effectiveness. Using this knowledge, multi-class linear regression (MCLR) [13], based on the FoCal toolkit [14], was used to combine individual sub-system scores.

5. COMPUTATIONAL RESOURCES

The UTD-CRSS LRE system was implemented on our high-performance Rocks computing cluster running the CentOS 5.5 Linux distribution. The cluster comprises 4 compute-nodes each equipped with 2 six-core Intel Xeon 2.67 GHz CPU's, yielding 48 processing cores. A total of 4×24 GB RAM is available internally on the system.

6. PROCESSING SPEED

The most time consuming task of our primary system is training the total variability matrix for the i-vector extraction. It takes about 12 hours on average. The UBM training requires about 1:43 (h:m) on our development set. For a 30-second speech file, CPU time required for each separate tasks are as follows:

- Feature extraction = 5.7 s
- i-vector extraction = 1.44 s
- Classification = 0.24 ms

7. SUBMITTED SYSTEMS

7.1. CRSS Primary

This is the i-vector system using MFCC-SDC. (crss_primary_llr.out).

7.2. CRSS Contrast 1

This is a fusion of the following systems: (i) i-vector using MFCC-SDC, (ii) i-vector using MHEC-SDC, (iii) SVM-GSV using MHEC-SDC, (iv) PPRLM, and (v) Articulatory, and Prosody features based. (crss_contrast1_llr.out).

7.3. CRSS Contrast 2

This is an alternate fusion of the following systems: (i) i-vector using MFCC-SDC, (ii) i-vector using MHEC-SDC, (iii) SVM-GSV using MHEC-SDC, (iv) PPRLM, and (v) Articulatory, and Prosody features based. (crss_contrast2_llr.out).

7.4. CRSS Contrast 3

This is the i-vector system using MHEC-SDC. (crss_contrast3_llr.out).

7.5. CRSS Contrast 4

This is the SVM-GSV using MHEC-SDC. (crss_contrast4_llr.out).

7.6. CRSS Contrast 5

This is the i-vector system using MFCC-SDC features. No calibration is used. (crss_contrast5_llr.out).

8. CONCLUSION

In this paper we briefly described the UTD-CRSS overall solution and sub-systems for the NIST-LRE 2011 competition. In our primary system the i-vector framework was employed, and the standard MFCC-SDC acoustic features were used to represent the various language classes under investigation. Observing the room reverberation effects on some portion of the training data, we also explored the effectiveness of our recently proposed MHEC features for the LRE task. As our secondary systems we investigated PPRLM, articulatory feature based, and prosody based system. The multi-class linear regression from the FoCal toolkit was used for both calibration and fusion purposes. In future, we plan to provide more robustness to our front-end feature extraction by including the vocal tract length normalization (VTLN) and feature domain nuisance attribute projection (NAP) compensation modules.

9. ACKNOWLEDGEMENT

The authors would like to thank Dr. Yun Lei and Dr. Niko Brummer for their valuable comments and suggestions in our development. We especially would like to thank Dr. Pietro Laface for providing us with their LRE-2009 development data. Also we thank Dr. Weiqiang Zhang for providing data source of SBS radio and valuable discussion.

10. REFERENCES

[1] A. Sangwan, M. Mehrabani, J.H.L. Hansen, "Language Identification using a Combined Articulatory Prosody Framework," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4400–4403.

[2] M. Mehrabani, H. Boril, J.H.L. Hansen, "Dialect Distance Assessment Method based on Comparison of Pitch Pattern Statistical Models," in *Proc. IEEE ICASSP*, Dallas, TX, Mar. 2010, pp. 5158–5161.

[3] S.O. Sadjadi and J.H.L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE ICASSP*, Prague, Czech Republic, Apr. 2011, pp. 5448–5451.

[4] G.D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in iVectors space," in *Proc. INTERSPEECH*, Florence, Italy, Sept. 2011, pp. 861–864.

[5] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, R. Dehak, "Language recognition via Ivectors and dimensionality reduction," in *Proc. INTERSPEECH*, Florence, Italy, Sept. 2011, pp. 857–860.

[6] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verifcation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, pp. 788–798, May 2011.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 345–354, May 2005.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, pp 308–311, May 2006.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," [Online]. Available: `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

[10] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 31–44, Jan. 1996.

[11] P. Schwarz, "Phoneme Recognition Based on Long Temporal Context," Ph.D. Thesis, Brno University of Technology, Czech Republic, 2009.

[12] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 901–904.

[13] N. Brummer, "Measuring, Refining and Calibrating Speaker and Language Information Extracted from Speech," Ph.D. Thesis, University of Stellenbosch, South Africa, 2010.

[14] N. Brummer, FoCal Multi-class, [Online]. Available: `http://sites.google.com/site/nikobrummer/focalmulticlass`