Boğaziçi University NIST SRE10 Submission

Oğuz Yılmaz, Erinç Dikici, Murat Saraçlar

May 7, 2010

The NIST 2010 SRE submission of Bogazici University includes two different systems for the core test (core-core condition). The first system uses a GMM/UBM methodology whereas the second one uses the SVM-based GMM supervector approach. The following sections briefly explain these two systems and the computational resources used.

1 System 1: GMM/UBM Baseline

We apply a GMM/UBM classifier for speaker verification. The submission file is named "BOUN_1_core_core_alternate_llr".

Software

Our GMM/UBM implementation is based on the BioSecure Reference System BECARS/HTK [1], which utilizes HTK [2] for feature extraction, UNIANAL [3] for pitch, energy determination and voice activity detection, and BECARS [4] for GMM modeling and scoring.

Feature Extraction

We extract 34 dimensional cepstral coefficients (16 MFCC + 16 Δ + energy + Δ-energy) out of every 20ms of speech with 10ms overlaps. Voice activity detection is applied by bi-Gaussian modeling of the energy component. The cepstral vectors are normalized and frames corresponding to silence are deleted.

GMM/UBM Modeling

We train GMMs with 1024 components, having diagonal covariance. UBM is constructed using 27 hours of speech from SRE08 data. MAP adaptation is used to obtain speaker models. The relevance factor is chosen as 14. Only the means of the components are adapted; covariance and weights are copied from the UBM.

Scoring

Log-likelihood ratio scores are calculated for each test utterance. Final decisions on whether a test utterance is accepted or rejected are given by thresholding with an appropriate value which minimizes C_{Det} over the SRE08 test data.

System Resources

Processing times will be given as if the programs were run on a single CPU.

- CPU/RAM: Intel® Xeon® CPU E5450 @ 3.00GHz, 56 cores in total, max 2GB of RAM used
- Feature Extraction Runtime: 40h total for core-core training and test files
- Training (MAP Adaptation) Runtime: 960h
- Test Runtime: Total of 4888h

2 System 2: SVM Baseline

The system is an SVM baseline setup which uses GMM supervector features. We train GMMs with 2048 components for both male and female cases separately. Supervectors constructed using the stacked mean vectors of these components are used in the support vector machine as input features. The submission file is named "BOUN_2_core_core_primary_other".

Software

We use SPro [5] for feature extraction, FIR Echo Canceller [6] for echo cancellation, and SVMTorch [7] for SVM training and testing.

Feature Extraction

34 dimensional features are extracted, having 16 MFCC, 16 Δ, energy and the Δ-energy. The lower and higher cutoff frequencies are 300 and 3140 Hz, respectively. 20ms Hamming window is used with 10ms increments. ASR files are utilized to remove non-speech segments. Features are mean subtracted and normalized using a three second window.

UBM Training

A 2048 mixture GMM is trained on about 12-14 hours of speech data for male and female universal background models using SRE 2008 data. Only half of the short2 and short3, core train and test files, are utilized in UBM training. The other half is used for SVM's negative examples training.

SVM Negative Examples

1604 male negative examples and 2217 female negative examples are adapted from the UBM. These negative examples are used to train impostor speaker models.

Speaker Models

All models are adapted from the UBM using MAP adaptation. Relevance factor is taken as 16. Using these adapted models, speaker SVM models are trained using the supervector linear kernel. Negative examples are obtained through half of the SRE08 short2 and short3 data. In the SVM training we have 1 positive example for the speaker against 1604 negative examples for male and 2217 negative examples for female cases.

Scoring

SVM decision output values are calculated as final scores. The threshold value for t/f decisions is selected via an analysis over the SRE04 data.

System Resources

Note : Feature extraction time including echo cancellation are also included.

- CPU: Intel® Xeon® CPU E7320 @ 2.13GHz
- UBM: 173.6h for male UBM training, 1.9GB RAM using 12hours of speech data
  180.1h for female UBM training, 1.9GB RAM using 12.4hours of speech data
- SVM Negative Examples: 458.04h for all male and female negative example training, 1.0GB RAM
- Speaker Models: 664.3h for all male and female speaker model training, 3.0GB RAM
- Testing: 1623.52h for all male and female testing, 1.0GB RAM

3 System 3: GMM/UBM and SVM Fusion

The third system is a basic fusion strategy between the outputs of GMM/UBM and SVM systems. For both setups the score threshold is set to 0, and fusion scores are calculated by averaging the systems' scores. The hard decision returns true if both systems accept the utterance, if not, it returns false. This third submission file is named "BOUN_3_core_core_alternate_other".

References

[1] A. Mayoue, "Reference System Based on Speech Modality BECARS/HTK," Technical report, GET-INT, 2008.

[2] "HTK Speech Recognition Toolkit," http://htk.eng.cam.ac.uk/.

[3] "UNIANAL Universal Speech Analysis and Synthesis," http://speech.fit.vutbr.cz/files/software/unianal/unianal.tar.gz.

[4] C. Mokbel, H. Mokbel, R. Blouet, G. Aversano, "BECARS Library and Tools for Speaker Verification," 2008, http://www.tsi.enst.fr/becars/index.php.

[5] "SPro Speech Signal Processing Toolkit," http://www.irisa.fr/metiss/guig/spro/.

[6] "FIR Echo Canceller," http://www.isip.piconepress.com/projects/speech/software/legacy/fir_echo_canceller/.

[7] R. Collobert, S. Bengio, "SVMTorch: Support Vector Machines for Large-scale Regression Problems," Journal of Machine Learning Research, vol. 1, pp. 143160, 2001.