RUN system description for SRE-2010

Mitchell McLaren and David van Leeuwen

1 Introduction

This is the first participation of Radboud University Nijmegen (RUN) to the NIST Speaker Recognition Evaluation since a long time. Our goals for this participation have been

- to obtain a basic speaker recognition system infrastructure
- to gain experience in calibration for the new operating point
- to obtain a normalized $C_{\text{det}} < 1$ for some of the conditions.

It may seem that we have set our goals not too ambitious, this is partly due to our the limited resources in time and computing infrastructure during the development. Most of the development efforts have been in attempting to port the JFA system used in the RUN submission to the Evalita'09 evaluation to this NIST evaluation, however, these efforts appear to have been in vain. We were not able to churn any decent performance out of the JFA system, and have decided to only utilize a classical GMM-SVM NAP system and a more recently developed dotscoring system for this edition of the NIST evaluation series.

2 Systems

The RUN system is a fusion of four very similar systems. The sub-systems are formed by the Cartesian product of two spectral feature types (PLP and MFCC) and two classifiers (GMM linear scoring, a.k.a. dotscoring, and GMM-SVM supervector system). The systems are fairly similar in design and share most of the training data and use of training data for the various components of the systems. A most notable difference is the use of Z-norm for the dotscoring system and the background data for the SVM system, which are similar in function. Contrary to earlier evaluations we were involved in (RUN Evalita 2009, TNO NIST SRE-2008), all systems are completely conditioned on speaker sex in UBM, background, compensation and normalisation speech data.

2.1 SAD, preprocessing and feature extraction

At the front-end, microphone data (long and short interviews, phone calls recorded using a room microphone) is treated slightly differently from telephone

data. Microphone data is first sent through a Wiener filter [1]. Then, feature extraction is carried out for both microphone and telephone data, delta's are calculated and silence frames, with energy below 30 dB below the segment's maximum frame energy, are discarded. Two types of feature extraction are employed: a) 12 PLP + log energy cepstral features computed over 32 ms frames with 16 ms steps, and b) 13 MFCC features (incl C_0) computer over 30 ms with 10 ms step size. The PLP computation does not use RASTA processing and uses a MEL filter bank. For both feature types, short time Gaussianization [5] over a period of about 4 seconds is applied.

2.2 UBMs and supervectors and sufficient statistics

Five hundred and twelve[1]-component UBMs for both speaker sexes are trained using a collection of NIST SRE speakers, including both telephone and microphone data. Given a UBM, for each speech file used we computed both the MAP [4, 6] adapted mean supervector and the zeroth and first order sufficient statistics of the features given the UBM $n_j = \sum_t p(j|x_t)$ and $f_j = \sum_t (x_t - \mu_j) S_j p(j|x_t)$, where we used Niko Brümmers Guassian-mean centered and Gaussian-precision scaled features. These two computations are in fact very similar. The supervector formed by the shift in mean from the UBM to MAP adapted means is used for an SVM system, and the sufficient statistics for a dotscoring system.

2.3 Channel compensation

We have used two different sets of training data for channel compensation. One is based on 2004 telephone speech segments, and one is based on 2005–2006 microphone speech segments. Two sets of intersession transform matrices for dotscoring and SVM were trained using these segments, for both telephone and microphone. Each of these retained the 30 dimensions with highest variance. The two matrices were stacked and normalized using singular value decomposition. For the SVM system we used NAP [3], for the dotscoring a factor analysis approach as taken by SDV for SRE 2008, where compensation is applied directly to the sufficient statistics.

2.4 Classifiers

2.4.1 Dot-scoring

The dotscoring system directly used compensated sufficient statistics to score test segments on speaker models. These models were computed in the scoring script, using a not-so-relevant factor of 1 for female data and 8 for male data. T- and Z-norm score normalization were both applied using cohorts consisting of a combination of microphone speech from SRE 2005 and 2006 and telephone

[1] Writing style rules forbid us to start a sentence with the number 512

speech from SRE 2004. There was no overlap between the speakers in these cohorts.

2.4.2 SVM

The SVM system performed classification using GMM mean supervectors adapted from 512-component gender-dependent UBMs. This system was based on the libSVM package and the open-license Mistral speaker recognition toolkit from LIA, France. Background datasets consisted of utterances sourced from both microphone and telephone speech from the NIST SRE 2004-2006 corpora. Ensuring that impostor speech segments contained a minimum of 15 seconds of speech activity was found to aid performance. T-norm score normalization was also applied using the background dataset as the T-norm cohort. For this task, T-norm models were trained using a 'leave-one-out' approach. The application of Z-norm was found to benefit SVM classification to a negligible degree and was, therefore, not applied.

3 Calibration and Fusion

Calibration was carried out using Niko Brümmer's calibration tools. Specifically, the 'scal' score-to-likelihood ratio function was used, a sigmoid-like function characterized by 4 parameters (lower limit, higher limit, slope and shift). In fusion, each additional system has its own slope parameter. We used the SRE-2008 evaluation data for training the calibration parameters. The objective function for calibration is C_{llr} [2], where for the integration a prior distribution around the NIST prior 0.001 is used. In order to have sufficient non-target trials around the DCF point, we generated all relevant model-test segment combinations from the SRE-2008 data. Since both the dotscoring and SVM systems compute the full score matrix anyway (because a matrix-matrix multiplication can be carried out quite efficiently), no additional scoring is necessary. We created three different calibration classes, based on the microphone types in the trial: these are 'mic-mic,' 'mic-tel' and 'tel-tel,' and each gender was calibrated separately. We used only English-English SRE-2008 model-test trials for calibration.

4 Submission

We submitted only a single system, and only the core condition, which therefore is primary. Contrary to other years, we did not participate in other consortia, nor submitted multiple systems, lowering the chances of coincidentally performing well in one submission or condition.

We participated in the extended test as well, because we had seen some benefit in a large magnitude of non-target trials during calibration, even at our level of performance. The basic principle of scoring was the same as for the core condition; we computed scores of all models against all test segments, and then later selected the trials according to the extended index. Certain memory

limitations of some 32-bit programs caused further unspecified head-aches, as did failing NFS-servers and full discs. Finally, during a sanity check, veryfying that the trials from the original list produced the same scores as computed while processing the extended list, we found that there appeared some systematic noise in the scores due to an uncharted bug in the fusion script.[2] We have tried to fix this, resulting in an extended submission file not entirely consistent with the primary submission, but hopefully with less noise in the scores.

5 Computation Time

The RUN submission was run on a cluster of six Intel Core2duo machines each with between 2–4GB[3] of memory. Feature extraction proved to be the bottleneck in terms of computational efficiency for the RUN submission. At an average of one MFCC or PLP feature per minute, this consumed around 300 hours of CPU processing time per feature set. From these features, statistics and GMM mean supervectors were collected simultaneously—this process took approximately 20 hours per feature set. Training and classification using the dot scoring system was completed in 20 minutes on a single CPU, while the SVM configuration consumed approximately 2 hours for training and 2 hours for testing. Quite some time was spent on ferrying the score matrices across the Atlantic in order to be able to do the calibration/fusion in California. The fact that our local uplink switch was configured at 10 Mb/s did not help. Sending back the calibration parameters (42 floating point numbers in total) fortunately happened over full-bandwidth IP connections.

6 Acknowledgements

This work was supported in part by the EC Marie Curie ITN project "BBfor2," under REA contract 238803. We would like to thank Marijn Huijbregts for his help on bootstrapping the JFA system and numerous fixes to the computing infrastructure. We are indebted to Niko Brümmer for the use of the calibration tools, and for fusing in side-information into our limited brains. We would also like to thank ICSI whos computing infrastructure we used for the calibration, since that required both large memory machines and a license to the commercial counterpart of GNU Octave. The NIST SRE10 group provided some entertaining discussions, this year.

References

[1] A. Adami, L. Burget, S. Dupont, G. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-ICSI-OGI features for ASR. In *Proc. ICSLP*, 2002.

[2] This may be considered 'interaction with the data' in a strict interpretation of the rules

[3] for a discussion on multipliers of bytes, see `http://xkcd.com/394/`

[2] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20:230–275, 2006.

[3] William Campbell, Douglas Sturim, Douglas Reynolds, and Alex Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. ICASSP*, pages 97–100, Toulouse, 2006. IEEE.

[4] J. L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Speaker-independent continuous speech dictation. *Speech Comm.*, 15:21–37, 1994.

[5] Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey*, pages 213–218. Crete, Greece, 2001.

[6] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.