VAS system for EVALITA 2009 Speaker Identity Verification - Application Track Task

Benoît Fauve[1], Driss Matrouf[2], Jean-François Bonastre[2], and John Mason[3]

[1]ValidSoft Ltd, London, UK
benoit.fauve@validsoft.com
[2]LIA, Université d'Avignon et des Pays de Vaucluse, France
{driss.matrouf,jean-francois.bonastre}@univ-avignon.fr
[3]Speech and Image Research Group, Swansea University, UK
j.s.d.mason@swansea.ac.uk

Abstract. This paper presents the system submitted for the EVALITA 2009 Speaker Identity Verification Application Track collaboratively developed by ValidSoft UK, LIA (Université d'Avignon et des Pays de Vaucluse) and the Speech and Image Research Group (Swansea University) under the acronym VAS (Validsoft-Avignon-Swansea). This work is based on the Alize speaker verification toolkit, with the singularity of using only Evalita development data and no T or Z forms of score normalisation. The discussion of results focuses mainly on issues with threshold setting and actual detection cost.

Key words: Speaker verification, Evaluation.

1 Introduction

VAS submission to EVALITA 2009 speaker identity verification application track (SIVAP) task is based on work developed by the participants in [1, 2] and uses ALIZE speaker verification toolkit [3]. The approaches used are the same as the one used by Swansea Univesity (UWS) and LIA in NIST Speaker Recognition Evaluation (SRE) 2008. The main difference lies in the use of background data to model universal background models (UBM) and channel matrices in factor analysis (FA). All the background data comes from Evalita development set.

2 Submission

There are two different systems used in the VAS primary (and only) submission: one system tailored to TS1, and one to TS2.

- **for TS1**: GMM-UBM low feature order (33)
- **for TS2**: channel FA modelling and high feature order (50)

For TS1 (short testing) the choice is to use a conventional GMM-UBM and lower feature order, following our observations in [4]. It is acknowledged that this standard approach is not the current state of the art; however the work of Kenny and others on JFA (see results on short duration task in [5]) and associated approaches require large amounts of background data and techniques still under development in our labs for the specific case of short duration.

On the other hand with TS2 we have chosen to use a FA based system as we found that the development set of Evalita was large enough to train intersession matrices and benefit from the channel compensation approach developed in [1]. The latest approach has proven to provide good results without the need of Z or T score normalisation. This is the main motivation not to use cohort based score normalisation in our submission (only standard UBM-GMM score normalisation). Also, the small development set made it difficult to observe clear trends with cohort selection when score normalisation was assessed. For similar reasons SVM based system as proposed in [1] has not been used.

The next subsection provides more details on our submission. The reader should also refer to [1] and [2].

2.1 Frontend

The motivation to use different feature orders depending on the segment length comes from observations in [4]. Also UWS submission at NIST SRE 08 used a fusion between MFCC and LFCC based systems, with the observation that LFCC fitted better female speaker (hypothesised higher resolution in high frequencies). For Evalita each gender has its own type of feature, MFCC for male and LFCC for female. To sum up features used are:

- MFCC for male and LFCC for female:
- for TS1: feature order 33 (16 cep, 16 deltas , delta Energy)
- for TS2: feature order 50 (19 cep, 19 deltas , 11 deltas, delta Energy)

Frame detection is performed with a threshold on the energy component. Features are normalised with Cepstral mean subtraction (and 1-variance).

2.2 Modelling Approaches

The systems are gender dependent. UBMs have 512 components and are trained on the Evalita UBM development data. Channel matrices are trained on the same data and their rank is 40.

2.3 Scoring and Thresholds

As mentioned, only UBM-GMM score normalisation is used (ie no T or Z forms). To deal with the variations in score range and the given channel, the average and standard deviation of impostor accesses are determined on the development set for a given group for a trial with same gender, training and testing (length

and channel type) conditions, and used to simply 'calibrate' the corresponding system output. This is a very simple approach aimed at equalising the impostor score distributions and improving the EER when considering a mix of GSM and PSTN test conditions. Note, it is not a log likelihood calibration. Different thresholds have been set depending on the training (tc1 to tc6) and testing conditions (ts1-ts2 + channel detected in test files) and gender. A total of 48 thresholds is determined on the Evalita development set.

3 Results and discussion

Results of the VAS primary system are provided in Table 1 with the standard measures of equal error rate (EER), minimum of detection cost function (minDCF) and three actual detection cost functions (actDCF)[1]. 'actDCF submission' and 'actDCF bugfix' correspond to the DCF after threshold setting as described in the previous section. A post evaluation result 'actDCF P-G together' is also given. In this particular case a single threshold is set per testing condition (no distinction is made between PSTN and GSM test conditions), hence reducing the total number of thresholds from 48 down to 24.

Table 1. Performance of VAS primary system in terms of EER, minDCF and actual DCF, given overall and per training and testing conditions.

Training condition	Testing condition	EER(%)	minDCF	actDCF submission	actDCF bug fix	actDCF P-G together
all	all	9.20	0.279	1.496	0.356	0.292
TC1	TS1	14.16	0.370	2.244	0.442	0.445
TC2	TS1	16.08	0.390	2.721	0.635	0.529
TC3	TS1	11.29	0.280	1.960	0.329	0.311
TC4	TS1	11.72	0.316	2.119	0.386	0.363
TC5	TS1	9.37	0.245	2.030	0.412	0.429
TC6	TS1	7.39	0.179	1.536	0.253	0.217
TC1	TS2	7.25	0.229	1.338	0.344	0.246
TC2	TS2	7.80	0.226	1.167	0.323	0.288
TC3	TS2	5.39	0.164	0.817	0.297	0.173
TC4	TS2	4.86	0.148	0.791	0.239	0.160
TC5	TS2	3.43	0.110	0.735	0.313	0.202
TC6	TS2	2.45	0.074	0.494	0.298	0.140

The difference between the minDCF and 'actDCF bugfix' values is relatively important suggesting some difficulties in the threshold setting. Some explanation

[1] The original VAS submission included a software bug relating to the evaluation of minDCF: local variables referred to NIST rather than EVALITA parameters. This had a significant adverse influence. VAS wish to thank Niko for accepting this account and including results, suitably labelled, with the bug fixed. This bugfix does not change any DET-curves, or EER or minCDF.

can be found on Figure 1 where optimal thresholds are given for the male TS2 conditions, on development and evaluation sets.

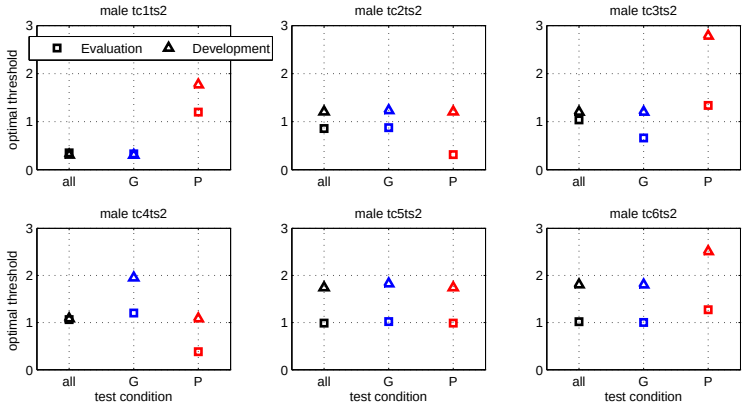

Fig. 1. Optimal threshold value on the male and TS2 subset on the development sets (triangles) and evaluation set (squares), on PSTN (P) and GSM (G) subset and without distinguishing between channels (all).

Except for a single case (tc1ts2 G) the optimal threshold on development set is systematically higher than the one on the evaluation set. Because VAS development process was quite general and did not involve any intensive parameter tuning to fit Evalita development data, we can hypothesised that the development set was not fully representative of the evaluation set and to some extent too optimistic. An alternative would have been to set thresholds on a smaller number of subsets (each training, testing, gender condition had its own threshold), to avoid thresholds to be too specific to the few trials of the given development subset. This is illustrated with the actDCF 'P-G together' values in the last column of Table 1; by not distinguishing between P-G tests results are closer to minDCF.

4 Conclusion

It should be noted that VAS submission have used only the data provided a Evalita development. It is to expect that better performance can be obtained through judicious use of other data sets, particularly for eigenchannel modelling, SVM and score normalisation. More work on this question is needed and it would fit well in a more general research theme on the a use of background data.

Evalita has proved beneficial in its focus on the Italian language and in providing an alternative to NIST.

References

1. Matrouf, D., Scheffer, N., Fauve, B., Bonastre, J.F.: A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification. In: Proceedings of Interspeech (2007)
2. Fauve, B., Matrouf, D., Scheffer, N., Bonastre, J.F., Mason, J.: State-of-the-Art Performance in Text-Independent Speaker Verification through Open-Source Software. IEEE Trans. on Audio, Speech and Language Processing, vol. 15, issue 7 (2007)
3. AIZE/MISTRAL platform website, http://mistral.univ-avignon.fr/en/
4. Fauve, B., Evans, N., Pearson, N., Bonastre, J.F., Mason, J.: Influence of Task Duration in Text-independent Speaker Verification. In: Proceedings of Interspeech (2007)
5. Kenny, P., Dehak, N., Ouellet, P., Gupta, V., Dumouchel, P.: Development of the primary CRIM system for the NIST 2008 speaker recognition evaluation. In: Proceedings of Interspeech (2008)