ValidSoft

ValidSoft - NIST 2010 SRE - System Description

Benoit Fauve

ValidSoft, 9 Devonshire Square, London EC2M, UK

benoit.fauve@validsoft.com

1. Systems Overview

ValidSoft systems for NIST SRE 2010 are based on support vector machine (SVM) with linear kernel applied to channel compensated Gaussian mixture models (GMM) supervectors [1][2]. Though this is the only approach used in the submission, different flavours of background data have been used to accommodate the different types of recordings in the core condition. To deal with the new DCF and the lack of false alarm examples in the development set, a method that estimates the decision threshold according to the means and variances of the target and impostor score distributions has been tested. It has been observed that this estimation of threshold is valid with old DCF parameters, but does not cope well with the new parameters. Alize LGPL toolkit [3] was used for Factor analysis modelling. Systems have been submitted on the core-core condition only.

2. Systems configuration

2.1. Frontend

All systems are based on cepstral coefficients (FFT frame size is 20 ms and the frame rate is 10 ms, bandwidth is limited to 300-3400Hz). 19 static coefficients are calculated out of 24 filter bands, to which are added 19 deltas, the delta energy and the 11 first double deltas resulting in features of size 50. Two simple types of speech activity detection (SAD) are used, both based on a tri-Gaussian modelling with EM of the energy component:

- Mean-based (M): the selection threshold is given by the mean of the Gaussian of highest energy (~45% of the frames are kept).
- Weight-based (W): the percentage of selected frames is given by the weight of the Gaussian component of highest energy (very selective ~25% of the frames are kept).

Each feature component is normalised by mean and variance estimated over the full segment length.

2.2. Speaker modelling and testing

All Gaussian mixture models have 512 components. With a feature of size 50 the resulting supervector size is 25 600. For each utterance a GMM with factor analysis based channel compensation [1] is estimated. SVM modelling and testing is carried out as described in [2]. When the channel matrix depends on two different types of recordings (as in VLD_2 subsystem 1, cf. table 1), two matrices are trained independently and concatenated.

2.3. Score normalisation

Two types of score normalisations have been used, T and TZ. When TZ is used the cohort for Z and T are the same. No score calibration has been used.

2.4. Development set and background data

NIST SRE 2008 short2-short3 and short3-long conditions have been used for system development. Using SVM and fast scoring some experiments have also been run on extended trial sets. Microphone speech from NIST SRE 2005 was found to be a good resource for channel compensation of interview segments. The balance of microphone and phone call

data in the UBM has been found after a series experiments on GMM-MAP systems with a low number of components (64) for rapid testing. For channel matrices the full set of phone call recordings from NIST SRE 04 and of microphone recordings from NIST SRE 05 have been used. For other background data the number of segments selected is given in Table 1. Of particular interest and when compared to systems presented in [4], some notable improvements come from using larger SVM cohorts.

2.5. Systems submitted

Full configuration details are given in the following tables:

Table 1.

Systems		Feature		UBM			Channel matrix	
		SAD	LFCC /MFCC	Set	# seg male	# seg female	Training Set	Rank
VLD_2	1	M	L	Phn04+Mic05	*354*	*352*	Phn04+Mic05	*80*
	2	M	M	Phn04	*219*	*196*	Phn04	*40*
	3	W	L	Phn04	*219*	*196*	Phn04	*40*
VLD_3		M	L	Phn04+Mic05	*354*	*352*	Mic05	*40*

Systems		Svm cohorts			Score normalisation cohorts				
		Set	# seg male	# seg female	Set	# seg male	# seg female	T	TZ
VLD_2	1	Phn04-05-06+Mic05	*981*	*1186*	Phn04+Mic05	*250*	*275*		x
	2	Phn04-05-06+Mic05	*981*	*1186*	Phn04	*115*	*119*	x	
	3	Phn04-05-06+Mic05	*981*	*1186*	Phn04	*115*	*119*	x	
VLD_3		Phn04+Mic05-06	*822*	*883*	Phn04+Mic05	*250*	*275*		x

The primary system VLD_1 is VLD_2 for phonecall-phonecall and phonecall-microphone trials and VLD_3 for microphone-microphone trials.

2.6. Decision threshold

In the hypothetical case when target and impostor score distributions follow Gaussian distributions (of mean and standard deviation m_T,s_T,m_F,s_F), the cost function can be expressed in terms of error functions (erf). Considering the cost function C as a function of the threshold θ, the equation $dC(\theta)/d\theta = 0$ take a quadratic form. A threshold θ_{opt} corresponding to the minimum of the DCF can be expressed with C_{miss}, C_{fa}, P_{target}, m_T, s_T, m_F and s_F. The idea here is to find the threshold according to estimates of the score distributions, which can be found without building a new protocol with a large number of impostor trials. Such an idea seems to be robust with former NIST's DCF parameters, but does not transpose well with the new parameters. Finally, when development results were available on extended development sets the chosen thresholds have been taken at minDCF. The other thresholds have been chosen according to the previously described $\theta_{opt}+bias$ (this bias was estimated from the results on extended data set where 'real' and 'theoretical' thresholds were available). Different thresholds were used according to the gender and channel type (microphone or phonecall).

3. Execution time estimates

The following estimates are given in a single thread mode on an Intel Core2 Q8300 @ 2.50GHz processor. In the case of a channel matrix of rank 40, the process from raw sound to GMM super vector vary between 0.0021 RT (selective SAD on conversational segments) and 0.0036 RT (less selective SAD on interview segments with greater density of speech). SVM training is done at a rate of 40 models per minute, but that could be drastically reduced by pre-computation of the GRAM matrix. SVM testing is a simple dot product. In the current implementation where models are not kept in memory the scoring of the 273787 trials of the male part of the core-core condition takes approximately 10 minutes (test segments already expressed as supervectors and without score normalisation).

4. Acknowledgements

ValidSoft would like to thanks the teams at LIA and UWS for the fruitful discussions, and the SRE10 Google group which made some steps in the development easier.

5. References

[1] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” in Proc. Interspeech, 2007

[2] W. M. Campbell, D. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Processing Letters, vol. 13, 2006.

[3] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, “ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition,” in Proc. Odyssey, 2008.

[4] B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, J. Mason "State-of-the-Art Performance in Text-Independent Speaker Verification through Open-Source Software" IEEE Transactions on Audio, Speech and Language Processing. Volume 15, Issue 7, Sept. 2007