Influence of task duration in text-independent speaker verification

Benoît Fauve[1], *Nicholas Evans*[2,1], *Neil Pearson*[1], *Jean-François Bonastre*[2], *John Mason*[1]

[1]Speech and Image Research Group, University of Wales Swansea, UK
[2]LIA, Université d'Avignon et des Pays de Vaucluse, France
{b.g.b.fauve.191992,n.r.pearson.161797,j.s.d.mason}@swan.ac.uk
{nicholas.evans,jean-francois.bonastre}@univ-avignon.fr

Abstract

Short duration tasks for text-independent speaker verification have received relatively little attention when compared to that directed at tasks involving many minutes of speech. In this paper we investigate verification performance on a range of durations from a few seconds to a few minutes. We begin with a state-of-the-art GMM-based system operating on a few minutes of speech per person and show that the same system is suboptimal on short (10 seconds) speech recordings. In particular we highlight that optimal frame selection exhibits a dependency on overall duration. This work sheds some light on the difficulties of transposing recent and important techniques such as SVM-NAP to the short duration tasks.

Index Terms: speaker verification, short duration, frontend optimization

1. Introduction

Over the last few years research emphasis in text-independent speaker verification has been directed predominantly towards long duration data tasks, driven perhaps by the initiatives to utilize higher-level information [1]. This means that typically a verification task is performed using at least a few minutes of speech and possibly much more, to train a speaker model and then to test it. The distribution of ever larger corpora, with numerous sessions available for each speaker from a large population has helped to validate techniques exploiting higher level information and has also contributed to research on session variability compensation. Techniques such as nuisance attribute projection (NAP) [2] and factor analysis (FA) [3] bring meaningful performance improvements by exploiting the prior knowledge on how speakers' characteristics vary over various recording sessions over a significant period of time. These new techniques have been successfully validated on relatively long duration tasks [4, 5]. It is acknowledged in [6] that short duration tasks have received relatively little attention (e.g. 10 seconds of speech for testing and training in the NIST evaluations [7]). However, these conditions are important in some commercial applications.

In this paper we show that state-of-the-art approaches that work well with long speech durations are not directly portable to short duration tasks, even when these systems do not employ higher-level information. We begin with a GMM system and its associated support vector machine GMM supervector linear kernel (SVM-GSL) with NAP compensation [4] giving state-of-the-art performance on speech intervals averaging 2.5 minute durations and demonstrate its sub-optimality on short speech recordings. We then report a series of experiments on short and long duration tasks where the percentage of dropped frames is varied and show a dependency between task duration and frame selection. The idea that different frames possess different discriminative powers is not new [8] and approaches such as frame weighting and dropping [9, 10, 11] aim to exploit this observation. Our experiments suggest that a frames discriminative power depends upon the context, i.e. the task duration. To support this idea we report further experiments, by considering various task lengths of between a few seconds and a few minutes and by using homogenous frames.

Utterance length compensation has already been applied in the context of likelihood score quality, [12] and [13] for example. Here we focus on tasks with given utterance lengths and highlight that different configurations lead to different raw performances according to the task length; consequently there is an argument for further session duration optimization in state-of-the-art speaker verification systems.

The remainder of the paper is organized as follows. In Section 2 we describe the experimental protocol and present our baseline results. An assessment of selective frame removal for one shorter and one longer duration task is reported in Section 3. Section 4 proposes an original set of experiments which guarantee homogeneity of frames over a range of task durations. We report some experiments which validate our results on the latest NIST'06 database in Section 5 before conclusions are drawn in Section 6.

2. Protocols and baseline performance

Initially we consider two speech conditions: (i) an average of 2.5 minutes for test and for train; (ii) 10 seconds for test and for train. They correspond to the 1conv4w-1conv4w and 10sec4w-10sec4w following NIST conventions which we refer to from now on as 1c1c and 10s10s respectively.

2.1. Systems configuration

Throughout the paper all background data (for UBM, Tnorm and NAP training) come from the NIST'04 database. We present results on the NIST'05 database and protocols. The NIST'06 database is reserved for final validation only.

The system is developed using SPRO[1] and ALIZE[2] open source toolkits, with the latest addition of SVM-GSL NAP . The core GMM is close to the description in [14], the only difference being standard feature optimization, with now 50 dimensions: 19 LFCC, their first-order derivatives, the first 11 second-order derivatives and the energy derivative.

Speech activity detection (SAD) is based on tri-Gaussian modeling of the energy distribution using an EM algorithm. The

[1]http://gforge.inria.fr/projects/spro
[2]http://www.lia.univ-avignon.fr/heberges/ALIZE/

threshold to determine those frames to discard is set to $\theta = \mu_3 - \alpha \times \sigma_3$ where μ_3 and σ_3 are the mean and standard deviation from the Gaussian of high energy, and α is a value controlling the selectivity. The higher α is, the larger the number of retained frames is. Initially a value of α=-0.5 is used which corresponds to an average of 28% of frames retained.

The UBMs are gender dependent with approximately 8 hours of 1side conversations from NIST'04 for each gender and have 512 components. Thus the mean supervectors have a dimension of 512*50=25600. The SVM-GSL models are trained using the world model samples as a cohort. For the 10s10s task a targeted cohort is used, i.e. 10s long segments. The NAP training is performed on the whole of the NIST'04 database. The NAP projection matrix has a rank of 40. Targeted T-norm is applied meaning that for each type of segment a similar length is used for the normalization cohort.

2.2. Performances

Figure 1 shows the performance of the described GMM system (GMM50) and the associated GSL-NAP system. The third profile (GMM33) comes from a GMM system found to perform better on the 10s10s condition. This second GMM-based system (GMM33) uses a different energy threshold (α=0.2) and only 33 feature coefficients (16LFCC+16Δ+ΔE)

There are two main observations from Figure 1:

- The difference in performance between the two GMM systems is surprisingly large. Results with GMM33 are noticeably better on 10s10s (27.3% equal error rate (EER) against 31.1% for the baseline) and close to the best reported performance at the NIST'05 evaluations.
- NAP compensation brings significant improvement on the 1c1c task (down to 6.19% EER from 9.22% for the GMM baseline), but it degrades results on 10s10s. This was also observed with our implementation of FA (not presented here). This research is still at an early stage, but there is a clear difficulty in porting these latest advances in speaker verification to short duration tasks.

The rest of this paper focuses on the first point above, investigating potential reasons which account for the differences in performance between the two GMM systems.

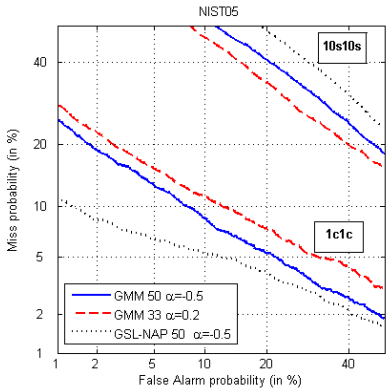

Figure 1: *DET curves for 10s10s and 1c1c NIST'05 datasets each with a GMM50, GMM33 and GSL-NAP50 configuration.*

3. Energy based frame selection

In this section we assess the impact of frame removal on GMM performance. We report a series of experiments with various values of α which controls the percentage of frames that are retained following speech activity detection. Table 1 gives the average number of frames kept according to α. Figure 2 shows the evolution in performance in terms of EER (top row) and minimum decision cost function (minDCF) (bottom row) as a function of alpha for a GMM system of feature order 33. It indicates that the 1c1c task gives better results (smallest EER and minDCF) with lower values of α in the order of -0.5 which corresponds to an average of 28% of retained frames. However, for the shorted duration 10s10s task the best results are obtained with values of α in the order of -0.1 which corresponds to an average of 35% of retained frames on a 1c sample. Of particular note is the striking difference in the EER and minDCF profiles for the two cases which show that the smaller values of alpha produce an increase in EER and minDCF for the shorter duration task. Thus a frontend that is optimized on a longer duration task will produce suboptimal results when tested on a shorter duration task. The reverse situation would appear not to be so severe in that a frontend optimized on a shorted duration task may still yield reasonable results when tested on a longer duration task.

Table 1: *Average percentage and average number of frames kept by the SAD on a range of energy parameters α on test segments from 1c1c and 10s10s NIST'05.*

	frame kept			
	10s		1c	
α	%	♯	%	♯
0.5	*48.6*	886	*49.4*	14812
0.2	*39.2*	714	*41.3*	12402
0	*33.9*	619	*37.4*	11224
-0.2	*29.6*	540	*33.7*	10102
-0.5	*24.0*	437	*27.8*	8348
-0.7	*20.0*	365	*23.4*	7011

As different UBMs are learnt for each value of α, it would seem that an acoustic space with a higher concentration on lower energy frames may offer greater discrimination when the quantity of speech is restricted. This observation has been confirmed by experiments which used a single UBM but where alpha was varied for training and testing. Results show this approach to be ineffective. Thus, in conclusion, acoustic unity between UBM and frames is important.

This is an interesting observation but limited to two different scenarios of 10 seconds and 2.5 minutes of speech. Also one of the differences between the 1c1c and 10s10s tasks lies in the way that the segments are obtained. The 1c sessions come from one side of a 5-minutes long conversation and so have a fixed length. The 10s segments, however, are portions from similar conversations but have a variable length as they are built with an energy-based detection method to retain an average of 10 seconds of speech. This results in a different density of speech from one type to another as suggested by NIST'05 ASR transcripts for test segments which indicate there is speech on an average of 35% of the 10s files against 41% for 1c files. This variation may result in an inconsistent selection of speech samples from our SAD.

To address the limitations of the experimental work de-

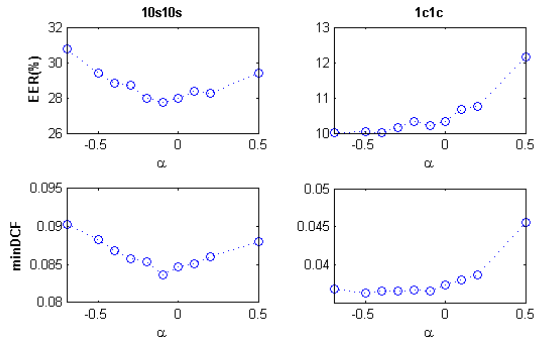

Figure 2: *EER(%) and minDCF for the 10s10s and 1c1c NIST05 data sets plotted against the energy threshold parameter, α.*

scribed above we now report further experiments which consider intermediate task durations and guarantee homogeneity of the frame selection procedure.

4. Intermediate duration with homogenous type of frames

Beginning with the 1c1c task we define a number of other variable duration tasks by selecting only a portion of the frames of speech found after energy-based thresholding.

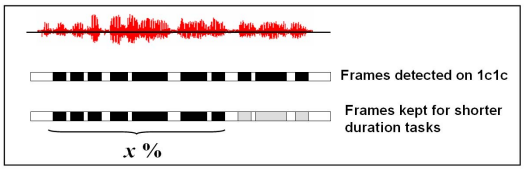

Figure 3: *Frame selection procedure for variable duration tasks. The percentage of retained frames, x, is varied simply by discarding the appropriate number of frames.*

Speech/non-speech detection is performed on all data from the 1c1c task as well as on the similar 1side segment from NIST'04 in order to maintain a targeted T-norm cohort. By keeping the first $x\%$ we have a task where the acoustic variety of frames is similar and where the average number of frames available is $x\%$ of the average number of frames available on the 1c1c condition. This gives a degree of consistency across the range of energy threshold.

The performance in terms of EER is illustrated in Figure 4 for six restricted tasks from 2.5% to 80%. The 100% value corresponds to the 1c1c condition itself. Three different GMM systems are tested:

- (GMM50 α=-0.5) and (GMM33 α=0.2) from Section 2 (with DETs on Figure 1).
- a third system transition between the two first with α=-0.5 and a feature dimension of 33. It only differs from (GMM50 α=-0.5) by the feature dimension and from (GMM33 α=0.2) by the SAD .

Table 1 shows that the number of frames selected for the 10s10s task is approximately 5% the number of frames selected for 1s1s, and reassuringly, the task restricted to 5% of the 1c1c frames gives similar results to the 10s10s condition (27.3% EER

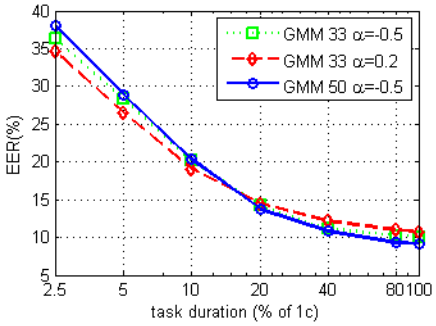

Figure 4: *performance in terms of EER(%) for various duration tasks specified in terms of percentage of kept frames.*

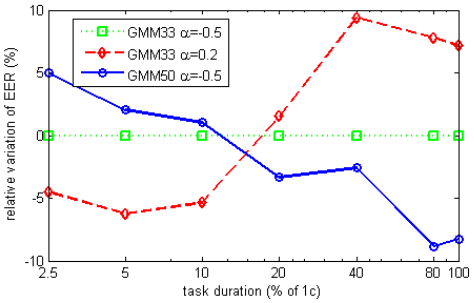

Figure 5: *Equal error rates, relative to the (GMM33 α=-0.5), for the (GMM50 α=-0.5) and (GMM33 α=0.2) configurations with varying percentages of data from the 1c1c condition of the NIST'05 database.*

in Figure 1 and 26.6% EER in Figure 4 for (GMM33 α=0.2)). A rough estimate of the equivalent quantity of speech for a $x\%$ limited task is $2 \times x$ seconds. Figure 4 provides an interesting profile of performance according to the amount of speech available. We see that from 20% to 100% of frames kept ($\sim$40s to 1c) the EER stays between 10% and 15%, whereas under $\sim$40s of speech the performances drop quickly.

As all the profiles are close on Figure 4 we plot on Figure 5 the variation in EER relative to the performance of the GMM33 α=-0.5 profile (hence this profile is flat in Figure 5). The profile corresponding to GMM30 alpha=0.2 (dashed profile) is less selective and corroborates the findings presented in Section 3 where such systems were better on 10s10s than on 1c1c.

The negative slope of the profile corresponding to GMM50 α=-0.5 in Figure 5 serves to illustrate that whilst double derivative features contribute positively to the longer duration tasks they have a negative contribution to the shorter duration tasks. The double differential process associated with the double delta features causes the resultant features to be noisy. The influence of this noise is greater when only smaller quantities of data are available. With longer speech durations the smoothing introduced by the GMM clustering process and the subsequent MAP process effectively attenuates this noise, to the extent that once the double derivative dimensions are reliably modeled they can contribute positively to classification. Some other interpretations of performance variations might be found in the GMM modeling itself, but this has not been investigated here.

On Figure 5, over all the restricted tasks the relative performance improvement (best system over GMM33 α=-0.5) is

in the range of 2.5% to 9%. This observation suggests that a speaker verification system operating with a potentially wide variation in task duration would benefit significantly by adapting its configuration according to the amount of data presented under testing conditions.

5. Results validation

The concept of a task dependant frontend has been successfully tested by the University of Wales Swansea (UWS) during the last NIST'06 evaluation campaign, obtaining good results for a GMM-only system in 10s10s and 1c1c tasks. We present here some results indicative of the performance.

Table 2: *performance on NIST'06 1c1c for systems: GMM50 2048 (UWS submission), GMM50 512 and SGL-SVM from GMM50 512*

	GMM50 2048	GMM50 512	SGL-SVM
EER (%)	9.11	9.43	6.44
minDCF	0.0354	0.0367	0.0199

Table 3: *performance on NIST'06 10s10s for systems: GMM50 512, UWS submission and GMM33 α=-0.1*

	GMM50 512	UWSsub	α=-0.1
EER(%)	30.1	**23.4**	25.1
minDCF	0.0904	**0.0832**	0.0855

Table 2 is related to results on the 1c1c task. The first column corresponds to UWS submission (SAD similar to α=-0.5). It uses 2048 Gaussians and provides state-of-the-art performance for a GMM-only system. The other 2 columns show the same system with only 512 Gaussians and its newly added GSL-SVM expansion. This last result assesses the goodness of the underlying GMM. Some results on the 10s10s task are given in Table 3. The first column corresponds to the previously mentioned GMM50 system found to be optimal on 1c1c. The second column is the UWS submission which was a fusion of 3 GMMs using different SADs and different feature dimensions. The third column is GMM33 with α=-0.1 which provides optimal performances in Section 3. The gain from the latest system is significant over GMM50 512. The UWS submission indicates there is room for further improvement in a GMM-only framework.

These various results support our findings on the development set.

6. Conclusions

In a similar manner to the way in which channel dependent UBMs have been used we suggest here an approach using different UBMs depending on the task duration. Extending the principal, in the same way that channel effects are now dealt with by employing sophisticated channel normalization processes, we could imagine some duration normalization inside the general speaker verification framework. Even if solutions at the score level are already used (the most popular being T-norm) additional approaches at the feature or model level may bring additional benefits. This paper also shows that some parts of speech with low energy might or might not contribute positively to the classification task depending on the quantity (and de-facto the quality) of speech available. The observed trends are to be confirmed on systems other than GMMs. The idea of using different UBMs, i.e. using different parts of the acoustic space to deal with duration issues is new. This finding can partially explain why NAP has proven difficult to transpose to short duration tasks as what appears as a nuisance in one context might contribute to discrimination in another. This aspect as well as the behavior of NAP and FA for short duration tasks requires further study.

7. References

[1] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *ICASSP*, 2003.

[2] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," in *ICASSP*, 2005.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *ICASSP*, 2005.

[4] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a gmm supervector kernel and nap variability compensation," in *ICASSP*, 2006.

[5] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *ICASSP*, 2006.

[6] M. Przybocki, A. Martin, and A. Le, "Nist speaker recognition evaluation chronicles part 2," Odyssey Workshop Presentation. [Online]. Available: www.speakerodyssey.com/templates/13.pdf

[7] A. Martin and M. Przybocki, "The NIST speaker recognition evaluation series, National Institute of Standards and Technology's website, http://www.nist.gov/speech/tests/spk."

[8] J. Eatock and J. S. Mason, "Phoneme performance in speaker recognition," in *ICSLP*, 1992.

[9] L. Besacier, J. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, pp. 89–106, 2000.

[10] J. Louradour, K. Daoudi, and R. Andr-Obrecht, "Discriminative power of transient frames in speaker recognition," in *ICASSP*, 2005.

[11] J. Pelecanos, D. Povey, and G. Ramaswamy, "Secondary classification for gmm based speaker recognition," in *ICASSP*, 2006.

[12] J. Pelecanos, U. Chaudhar, and G. Ramaswamy, "Compensation of utterance length for speaker verification," in *Odyssey*, 2004.

[13] W. Campbell, D. Reynolds, J. Campbell, and K. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *ICASSP*, 2005.

[14] J.-F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf, "Nist04 speaker recognition evaluation campaign: new lia speaker detection platform based on alize toolkit," June 2004, NIST SRE'04 Workshop. Toledo, Spain.