interspeech 2008 incorporating SST 08

Analysis of impostor tests with high scores in NIST-SRE context

Salah Eddine Mezaache[1,2], *Jean-François Bonastre*[1], *Driss Matrouf*[1]

Université d'Avignon, LIA
339 ch des Meinajaries, BP 1228, 84911 Avignon CEDEX 9, France
jean-francois.bonastre,driss.matrouf@univ-avignon.fr

[2]Centre Universitaire de Bordj-Bou Arréridj, Algérie
mez.salah@gmail.com

10.21437/Interspeech.2008-150

Abstract

In speaker recognition, performance of a system is usually estimated globally on a large set of tests, even if it is well known that some subsets of tests could show a very different behavior from the complete set. In fact, a small subset of tests could represent the main part of the reported errors. In this work, we highlight a such subset of tests, for which impostors obtain very high recognition scores. We evaluate if the problem comes from the involved speakers, from the voice excerpts or from the client model estimation technique. We also propose a strategy in order to minimize the effects of the observed phenomena on the overall performance of the system.

1. Introduction

During the past years, performance of text independent speaker recognition systems increases significantly as illustrated by NIST-SRE results[1]. EERs obtained with a single system are lower than 5%. This progress is mainly due to inter-session variability reduction techniques like Latent Factor Analysis (LFA) [5]. In this context, it becomes interesting to analyze more deeply system performance, usually described by the global EER or DCF, since a large difference between two EERs could come from a small number of trials or from a small number of speakers.
This paper proposes an analysis of scores issued from a state-of-the-art system applied on the NIST-SRE 05 and 06 corpora. It focuses on impostor trials for two main reasons. Firstly, there is about ten time more impostor trials than target trials in a classical NIST-SRE protocol, helping a statistical analysis. Secondly, we note that a non negligible number of impostor trials obtains a very high score, sometimes higher than a good target trial.
Section 2 presents the experimental context of this work (system, protocols and data). Section 3 is dedicated to the core of this paper : a detailed analysis of impostor scores. Section 4 proposes to take advantage of the previous section in order to improve the robustness of a speaker recognition system. Finally, the last section presents some conclusions and future works.

[1] http ://www.nist.gov/speech/tests/spk/

2. Experimental context

The experimental part of this paper is based on the NIST-SRE 2005 and 2006 evaluations, core task, all tests (*core-test 1conv4w-1conv4w, all, det 1*). The audio segments are issued from telephone conversations and have a duration of about 2,5 minutes of speech. The performance evaluation is done using the classical EER and minDCF criteria. Table 1 presents details of the NIST-SRE05 and 06 experimental protocols.
All the experiments are done using the open source ALIZE/SpkDet system developed by the LIA[2][3, 2]. ALIZE/SpkDet follows the classical UBM/GMM approach[7]. The configuration used in this paper, GMM-UBM-LFA (GMM-UBM with the Symmetrical Factor Analysis [4]), is detailed in [1].

Tab. 1 – Experimental protocols for the NIST-SRE05 and NIST-SRE06 *det1* (all tests)

	05		06	
Gender	male	female	male	female
Client models	274	372	354	462
Target tests	1231	1535	1570	2042
Non target tests	12317	16109	20561	27275
Total tests	13547	17643	22130	29316

3. Result analysis

Table 2 presents performance of the baseline system depending on the score normalization technique, for NIST-SRE05 and NIST-SRE06.

Figure 1 shows the target and non target score distributions for the female part of NIST-SRE06, using ZTnorm (distributions are very similar for the other test sets and score normalization techniques). It appears clearly that a significant number of impostor tests obtains a high score, comparable or higher than the mean of the target scores. We highlight this part of the impostor score distribution in this figure by a rectangle. The two vertical lines show the EER (left line) and the minDCF (right line) thresholds .
If it is quite normal to observe some impostor tests

[2]http ://alize.univ-avignon.fr/

TAB. 2 – DCFmin(x100) and EER% for the baseline system on NIST-SRE05 and NIST-SRE06

	05				06			
	male		female		male		female	
	DCFmin	EER	DCFmin	EER	DCFmin	EER	DCFmin	EER
nonorm	1.84	4.17	2.99	7.22	2.25	5.08	2.58	6.13
Tnorm	1.80	4.33	2.65	6.90	2.03	4.64	2.36	5.83
Znorm	1.66	4.08	2.49	7.49	2.16	4.32	2.54	5.84
ZTnorm	1.65	4.49	1.88	7.10	2.06	3.95	2.12	4.7

with a score higher than the DCFmin threshold, these scores are expected to remain only slightly higher than this threshold. It is not the case in the presented score distribution, where about two percents of the impostor scores are equal or higher than the average of the client scores. The next part of this paper is dedicated to this subset of tests, denoted here "problematic impostor tests". The "problematic" subset is composed of impostor tests which obtain a score higher than the DCFmin threshold (a posteriori determined). In the presented example, the "problematic" subset is composed of 217 tests. Figure 2 shows the distribution of the "problematic impostor" scores compared to the distribution of the client scores.

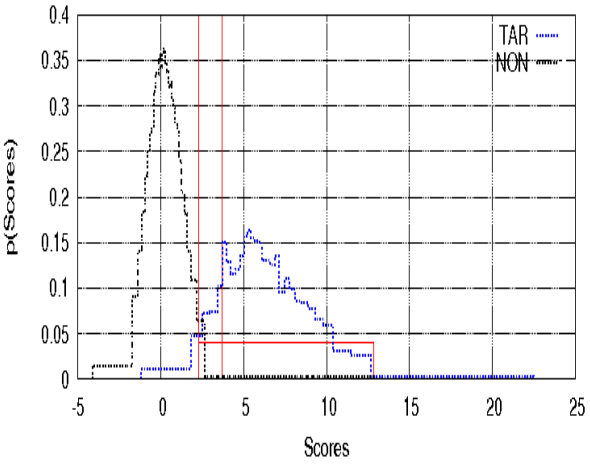

FIG. 1 – Target and non target score distributions on NIST-SRE06, female, ZTnorm.

An "problematic" impostor test is composed of a couple (seg_{train},seg_{test}), where seg_{train} is the segment used to train the corresponding target model and seg_{test} is the test segment. Of course, the couple of segments are pronounced by different speakers (we focus on non target tests). Table 3 shows the number of tests in the "problematic" subset, depending on the score normalization

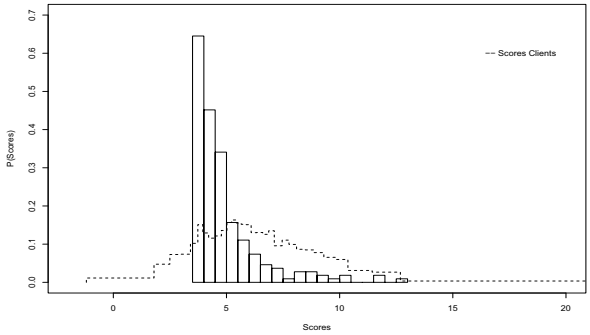

FIG. 2 – Score distribution of the "problematic impostor tests" subset and of the target scores, on NIST-SRE06, female, using TNORM

TAB. 3 – Details on the "problematic" subset on NIST-SRE06, female

	# tests	# seg_{train}	# seg_{test}
nonorm	154	76	143
tnorm	207	94	183
znorm	265	129	223
ztnorm	179	94	162

used, using NIST-SRE06, female. It details also the number of different seg_{train} and seg_{test} involved in the subset. In the selected protocol, 29316 tests are done, 27275 are non target tests and the cardinal of the "problematic" subset varies between 190 and 300, about 1% of the total number of impostor tests. It is interesting to notice that a quite large number of training and test segments are involved in this "problematic" subset (194 seg_{test} and 99 seg_{train} for ZTnorm). Regarding this observation, it seems that the high impostor scores are not linked to a specific subset of speakers, i.e. to some intrinsic speaker voice characteristics, as a quite large number of speakers are represented in this subset. The subset is quite independent of the score normalization technique, as 113 identical couples are in the subset for all the score normalization options (involving 50 seg_{train} and 105 seg_{test}).

In order to evaluate the influence of this "problematic" subset of impostor trials on the overall performance of the system, we propose in figure 3 the Det curves of the baseline system and of an "oracle" experiment where the tests belonging to the "problematic" subset are withdrawn (a posteriori). Table 4 presents the same results in terms of DCFmin for NIST-SRE05 and NIST-SRE06 (using ZTnorm). When less than 1% of the impostor tests are withdrawn a relative improvement of the DCF between 20% and 40% is noticed.

A validation experiment was also ran, by selecting randomly several subsets of impostor tests (each selected subset has the same number of tests than the "problematic" subset). No difference was noticed, compared to the baseline system. This result confirms the fact that the

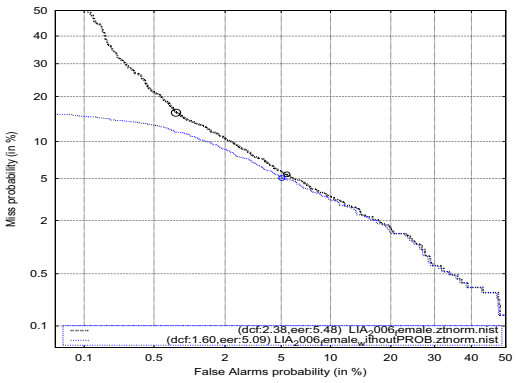

FIG. 3 – DET of the baseline system and the "oracle" system (without the "problematic" tests) for NIST-SRE06, Female, ZTNORM.

TAB. 4 – DCFmin (x100) of the baseline system and the "oracle system"

05		nonorm	tnorm	znorm	ztnorm
Fem.	Baseline	2.99	2.65	2.49	1.88
	Oracle	2.03	1.62	2.02	1.34
	gain(%)	**32**	**39**	**19**	**29**
Male	Baseline	1.84	1.80	1.66	1.65
	Oracle	1.13	1.32	1.04	1.15
	gain(%)	**37.5**	**26.6**	**37.3**	**30.3**
06		nonorm	tnorm	znorm	ztnorm
Fem.	Baseline	2.58	2.36	2.54	2.12
	Oracle	2.01	1.60	1.56	1.46
	gain(%)	**22.1**	**32.2**	**38.6**	**31.1**
Male	Baseline	2.25	2.03	2.16	2.06
	Oracle	1.64	1.32	1.33	1.40
	gain(%)	**27.1**	**35**	**38.4**	**22.3**

"problematic impostor tests" have a specific behavior. Several hypotheses could be proposed to explain the specific nature of these impostor tests with an amazing high score. As the underlined factors are complex to identify, we propose to explore several simple factors in the next subsections.

3.1. Amount of data

The specific nature of the "problematic tests" could be the amount of data, i.e. the length of the related speech segments. Table 5 shows some statistical information concerning the length of the segments. Even if a small number of training segments presents a short duration, there is no general statistical difference between "normal" and "problematic" segments in terms of duration.

3.2. Phonetic content of the segments

The amount of phonetic information present in the related segments is interesting to explore. Unfortunately,

TAB. 5 – Average duration for "problematic" speech segments, compared to "normal" segments

	05		06	
Gender	male	female	male	female
Prob. test seg.	7640	7450	8100	7920
Other test seg.	7850	7730	7800	7900
Prob. train seg.	7690	7890	7900	8159
Other train seg.	8015	7820	7900	8000

it is a complex issue as the links between the phonetic content and speaker specific information (taken into account by the GMM-UBM approach) is not trivial and quite unknown. We just listen the segments. As for the duration factor, few segments (usually the shorter ones) have a very small phonetic richness (short answers like enh, oueh, yah,...) but this potential factor can not be generalized to all the problematic segments.

3.3. UBM or client model behavior

Another factor could be an underlined limit of the GMM-UBM statistical approach. A model could output unexpected likelihoods inside the GMM-UBM paradigm when training or testing data are unusual, for example with a low variability. We looked at the likelihood distributions for both the normal (target and impostor) and "problematic" test segments given the UBM and a client model. No statistical difference were noticed.

3.4. Speaker specificities

As shown by [8] some speakers seem to have specific abilities to impersonate or to be impersonated by other speakers.
In order to evaluate the ability of a given speaker to impersonate a given speaker, for each "problematic" couple (Seg_{train},Seg_{test}), we compute all the tests corresponding to the available segments belonging to the same speaker than Seg_{test}, using the same model, issued from Seg_{train}. We do that only when at least three test segments are available for the test speaker. 68 couples of speakers are selected for this experiment, using this "at least three" rule, involving 408 tests in total and 63 different target models. For only one couple, all the corresponding trials (5) fall into the "problematic" class, indicating that the impostor is very close to the corresponding target speaker. For two other couples, the ratio of "problematic" tests is slightly over 50% (3 tests, including the original "problematic" test, among 5 are "problematic"). In total, 105 tests - to be compared to the 408 tests done in this experiment - are "problematic" (including the 68 original tests). Less than 11% of the additional tests are "problematic" even if the selected couples for this experiment are expected to present a similarity between the involved speakers.
The next experiment consists in scoring the test segments, Seg_{test}, involved in the "problematic" test set on a large set of speaker models. The experiment is done using NIST-SRE05 and 1348 speaker models (these 1348 models correspond to few hundreds of different spea-

Tab. 6 – Number of "problematic" trials for backward oracle and baseline systems, ZTnorm

System	05		06	
	male	female	male	female
Baseline	64	89	137	179
Back.Oracle	13	20	55	85

kers, as the database is composed of recordings from few hundreds of speakers only). Unfortunately, this impostor score distribution is very close to the classical one, we don't observed any "problematic" scores in this case. It seems that a test segment is "problematic" for a given speaker but not for other speakers.

3.5. Backward scoring

In this sub section, we evaluate if a classical backward scoring, where a model is trained using the test segment and compared to the target speaker training segment impacts on the "problematic" impostor subset. Table 6 presents the number of "problematic" couples for the forward system (baseline) and for the backward scoring applied only on the "problematic" subset of impostor trials (backward/oracle). A large decrease in the number of "problematic" trials - the number of "problematic" tests is divided between 2 or 3 times - is observed using the backward scoring in this "oracle" mode. It indicates that the high impostor score problem is more linked to the manner to handle a given speech segment than to the intrinsic properties of a given speaker or a given speech segment.

4. System robustness

As seen in the previous section, the backward scoring seems able to solve a large part of the "problematic" test problem. It seems interesting to use this scoring in order to improve the overall performance of the system. Table 7 presents the DCFmin for the baseline, the baseline without the "problematic" tests (NO-PROB), an oracle mix between forward and backward scoring where only the "problematic" tests are scored using the backward method (MIX) and an arithmetic fusion involving forward and backward methods (Fusion). As expected, the oracle mix shows an interesting gain in DCF compared to the baseline. When this Mix system is compared with the NO-PROB system, where the "problematic" tests are simply removed, it appears that about 40% of the loss due to the "problematic" tests is withdrawn. The arithmetic fusion between the two scoring methods (fusion system) allows a small gain compared to the baseline but is clearly less efficient than the oracle mix system. It indicates that a more complex fusion strategy is needed to improve the overall performance of the system.

5. Conclusion

This paper is based on the NIST-SRE framework, where presented systems showed a huge gain during the past years, in terms of EER and DCF. We started with a state-of-the-art system based on the GMM-UBM paradigm associated with a Latent Factor Analysis based inter-session mismatch reduction technique.

Tab. 7 – DCFmin for the baseline, backward,NO-PRO,MIX and Fusion systems (all with ZTNORM).

System	05 male	05 female	06 male	06 female
Baseline	1.65	1.88	2.06	2.12
Backward	1.79	1.89	2.00	2.15
NO-PROB	1.15	1.34	1.40	1.46
MIX	1.46	1.67	1.78	1.84
Fusion	1.67	1.77	1.92	2.08

We focused on a small subset of impostor trials, less than 1% of the total number of tests, which are responsible for about half of the system errors. After an analysis of this test subset, it seems that two factors play an important role : few speakers are very close one to the other in the view of such a system, and the system is not always able to train correctly a speaker model using a given speech excerpts, even if this recording embeds enough speaker specific information. We showed that different strategies based on a forward/backward scoring are able to decrease the influence of the latter factor on the system performance.

The results presented in this paper should be confirmed by increasing drastically both the size of the database and the number of tests, which seems possible by polling all the NIST-SRE databases. It seems also interesting to look at the correlations between the phonetic content of speech recordings and the two factors highlighted in this paper.

6. References

[1] Fauve et al., "State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software", IEEE Trans. Audio, Speech and Language Processing.,5(7) :1960–1968, 2007.

[2] Bonastre et al., "ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition", Speaker Odyssey, South Africa, January 2008.

[3] Bonastre et al., "ALIZE, a free toolkit for speaker recognition", Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA.

[4] Matrouf et al., "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification", INTERSPEECH Conference, Antwerp, Belgium,2007.

[5] Kenny et al., "Factor Analysis Simplified", Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA.

[6] Reynolds, D. A., "Speaker identification and verification using gaussian mixture speaker models", Speech Communication.,17(1-2) :91–108, 1995.

[7] Bimbot et al., "A tutorial on text-independent speaker verification", EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing, 2004.

[8] Doddington et al., "SHEEP, GOATS, LAMBS and WOLVES : a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation", Proceedings of International Conference on Spoken Language Processing (ICSLP 98), 1998.