Forensic speaker recognition

[Joseph P. Campbell, Wade Shen,
William M. Campbell, Reva Schwartz,
Jean-François Bonastre, and Driss Matrouf]

Forensic Speaker Recognition

[A need for caution]

© BRAND X PICTURES

There has long been a desire to be able to identify a person on the basis of his or her voice. For many years, judges, lawyers, detectives, and law enforcement agencies have wanted to use forensic voice authentication to investigate a suspect or to confirm a judgment of guilt or innocence [3] [35]. Challenges, realities, and cautions regarding the use of speaker recognition applied to forensic-quality samples are presented.

Identifying a voice using forensic-quality samples is generally a challenging task for automatic, semiautomatic, and human-based methods. The speech samples being compared may be recorded in different situations; e.g., one sample could be a yelling over the telephone, whereas the other might be a whisper in an interview room. A speaker could be disguising his or her voice, ill, or under the influence of drugs, alcohol, or stress in one or more of the samples. The speech samples will most likely contain noise, may be very short, and may not contain enough relevant speech material for comparative purposes. Each of these variables, in addition to the known variability of speech in general, makes reliable discrimination of speakers a complicated and daunting task.

Although the scientific basis of authentication of a person by using his or her voice has been questioned by researchers (e.g., by scientists in 1970 [4], British academic phoneticians in 1983 [5], and the French speech communication community from 1990 to today [6]), there is a perception among the

Digital Object Identifier 10.1109/MSP.2008.931100

1053-5888/09/$25.00©2009IEEE

general public that it is a straightforward task. As shown in [6], this misunderstanding partially began in 1962 with an article by Kersta, which appeared in *Nature* [7]. The article introduced the misleading term, voiceprint identification, which is still in vogue in daily newspapers, televised police dramas, and spy films. This term, voiceprint, leads many people to falsely believe that a graphical representation of the voice, via a spectrogram, is just as reliable as the structure of the ridges and minutiae of the fingertips or genetic fingerprints (e.g., DNA) and that it allows for a highly reliable identification of the original speaker. This misconception complicates the work of those in the forensic speaker recognition domain by introducing a false premise that all voices are unique, and discernibly so, under most conditions. Combating this mindset has become an ongoing process [30].

DURING THE LAST DECADE, A LARGE PART OF THE EFFORT DEDICATED TO THE SPEAKER RECOGNITION FIELD CONCERNED THE MISMATCH BETWEEN THE TRAINING AND TESTING SESSIONS.

No two speakers are absolutely the same, differing somewhat in anatomy, physiology, and acoustically. Even identical twins can have similar acoustics and differ in their implementation of a single segment in their linguistic system [31].

In forensics, it is not sufficient to state how similar two speakers are, and typicality must also be addressed. To do this, an examiner compares evaluation parameters of the speaker at hand to a larger reference sample of speakers. A measure of typicality helps quantify the strength of the forensic evidence, which is presented in the form of a likelihood ratio of two probabilities. Automatic speaker recognition systems can aid the forensic examiner in estimating the likelihood ratio.

With the developments in automatic speaker recognition over the last decade (e.g., [8] and [9]), there is an increased need to distinguish between its appropriate and inappropriate uses in various forensic voice authentication contexts and to differentiate between common versus forensic speaker recognition applications. In 2003, several scientific institutions reported on the status of the use of automatic speaker recognition technologies in the forensic field [2]. They concluded by sending a clear need-for-caution message, including statements such as, "currently, it is not possible to completely determine whether the similarity between two recordings is due to the speaker or to other factors...," "caution and judgment must be exercised when applying speaker recognition techniques, whether human or automatic...," or "at the present time, there is no scientific process that enables one to uniquely characterize a person's voice or to identify with absolute certainty an individual from his or her voice."

After these conclusions, the progress observed in the speaker recognition area has been very impressive, as shown in the National Institute of Standards and Technology (NIST) evaluation campaigns [10], [32], [33]. The major advancement was the appearance of new session variability modeling techniques, like the latent factor analysis (FA) or nuisance attribute projection (NAP) [11], [16]. The resulting level of performance encourages the use of automatic speaker recognition techniques in the forensic field. This article aims to comment on this progress and to evaluate if the 2003 need-for-caution message should be changed.

This article presents a summary of the progress made in the automatic speaker recognition field during the last few years and addresses the pertinence of the progress based on error rate criterion. The experimental context of this article, based on the NIST speaker recognition evaluation (SRE) evaluations, is described in the next section. The statistical Gaussian mixture model universal background model (GMM-UBM) approach used in the majority of the state-of-the-art systems is discussed in the following section, and the progress realized during the past years is also detailed. Evaluation of the performance and orientation of the research based on the objective of reducing error rates are presented. Finally, our views for future research and a conclusion with our message of caution are presented.

THE NIST-SRE FRAMEWORK

The NIST-SRE began in 1996. Since then, NIST organized an SRE campaign annually, with few exceptions. The main objective of the NIST-SRE is to provide an integrated framework for scientifically evaluating the approaches and systems in the field of speaker recognition: the participants work on the same corpus and protocols, the same performance criterion, and are time-synchronized by the campaign schedule. The main interest for the participants is the availability of free, large, specifically designed speech corpora that are enlarged or renewed each year. The NIST evaluations are mainly funded by the U.S. Department of Defense, which has a double objective. First, the campaigns are a showcase for the highest-performing speaker recognition techniques and serve as a good place to select the most promising research directions. Second, the sponsor has an important impact on the focus of the research done by all the participants, by proposing new tasks or protocol evolutions. The success of the NIST-SRE is confirmed year after year, as shown both by the growing number of participating sites (more than 40 in 2008) and by the number of NIST-SRE-related scientific publications in major conferences and journals.

NIST-SREs involve a text-independent speaker recognition task, mainly based on telephonic conversational speech. The systems have to answer the question, "did speaker X produce the speech recording Y and to what degree?" In this article, we focus on NIST-SRE core task. All the speech records are extracted from two-speaker telephonic conversations of about 5 min in duration. Only one channel is kept, giving on average 2¼ min of speech per recording.

THE GMM-UBM APPROACH

The GMM-UBM approach is the dominant one in text-independent speaker recognition [13]. This approach is based on a statistical modeling paradigm, where a hypothesis is modeled by a GMM model:

$$p(x|\lambda) = \sum_{i=0}^{i<m} \alpha_i N(x|\mu_i, \Sigma_i), \qquad (1)$$

where α_i, μ_i and Σ_i are, respectively, the weights, the mean vectors, and the covariance matrices (generally diagonal) of the mixture components. During a test, the system has to determine whether the recording Y was pronounced by a given speaker S. This question is modeled by the likelihood ratio:

$$\frac{p(y|\lambda_{\text{hyp}})}{p(y|\lambda_{\overline{\text{hyp}}})} \geq \tau, \qquad (2)$$

where Y is the test speech recording, λ_{hyp} is the model of the hypothesis where S pronounced Y, $\lambda_{\overline{\text{hyp}}}$ corresponds to the model of the negated hypothesis (S did not pronounce Y), $p(y|m)$ is the GMM likelihood function, and τ is the decision threshold. The model $\lambda_{\overline{\text{hyp}}}$ is a generic background model, the so-called UBM, and is usually trained during the development phase using a large set of recordings coming from a large set of speakers. The model λ_{hyp} is trained using a speech record obtained from the speaker S. It is generally derived from the UBM by moving only the mean parameters of the UBM, using a Bayesian adaptation function.

[FORENSIC APPLICATIONS OF SPEAKER RECOGNITION SHOULD STILL BE TAKEN UNDER A NECESSARY NEED FOR CAUTION.]

TRACING THE PERFORMANCE EVOLUTION DURING THE PAST YEARS

In this article, we limit reporting to the period 2004–2008, because the task remained constant during that time and general progress was observed. Our performance report is based on the work done in the Laboratoire Informatique d'Avignon (LIA), Université d'Avignon et des Pays de Vaucluse (different works were realized thanks to several cooperations, mainly with the Swansea University), a global representative of the general evolution of the speaker recognition domain. All the presented systems and approaches are integrated in the freely available open source system ALIZE/SpkDet (ALIZE/Technolangue http://www.technolangue.net/ and MISTRAL/RNTL http://mistral.univ-avignon.fr/) [14].

The starting point of this performance report is the LIA-04 system. This system, presented by the LIA during the 2004 SRE campaign, obtained state-of-the-art performance for a cepstral GMM-UBM system. It was slightly optimized for the 2006 campaign, mainly at the feature extraction level (for example, by increasing the feature vector dimension to 50), producing a relative decrease of about 10% of the equal error rate (EER). The EER is the point at which the probability of false alarm (false acceptance) is equal to probability of a miss (false rejection). This optimized system will be used as the reference baseline system in this article. The main results presented are issued from [15], where all the configuration details are given.

THE MIXED GMM AND SVM APPROACH

The discriminant classifiers based on support vector machines (SVM) were of great interest in the speech field. In speaker recognition, an important evolution was proposed, mainly by [12]. It uses a mixed approach, associating the robustness of the statistical modeling provided by the GMM-UBM paradigm with the discriminant power of the SVMs. This approach, denoted GMM supervector SVM with linear kernel (GSL), uses the GMM-UBM to model the training or testing data. Each recording is summarized by a supervector extracted from the corresponding GMM (obtained from the UBM by the maximum a posteriori procedure), composed by the concatenation of the mean coefficients of all the GMM components. The supervectors are then used as inputs of the SVM classifier (with a linear kernel).

Table 1 shows the performances of such GSL systems compared with GMM-UBM systems (and with the GMM-UBM 2004 system, LIA-04). The relative gain between the GSL system and the baseline (GMM-UBM) is about 18% in terms of EER and about 14% in terms of the minimum detection cost function (minDCF). The minDCF is a value of the detection cost function, which is defined as the weighted sum of the miss and false alarm error probabilities, using an ideal threshold. The parameters of this cost function are the relative costs of detection errors and the a priori probability of the target [36]. (The actual DCF does not assume the use of an ideal threshold.)

DEALING WITH THE SESSION MISMATCH

During the last decade, a large part of the effort dedicated to the speaker recognition field concerned the mismatch between the training and testing sessions. This mismatch comes from all the variability factors between two different recordings, except the interspeaker variability: environment, microphone or handset, transmission channel; psychological and pathological state of the speaker, linguistic content, voice aging, etc. The different works proposed in the literature mainly concern the first few factors, because of the influence of the NIST-SRE and its sponsor. Several solutions were proposed to deal with this intersession mismatch (with some important gains) at the acoustic level [17], [18] or the score level [19].

More recently, a new class of approaches was initiated by the works on the FA proposed by [11] for the GMM statistical paradigm and in [20] in the framework of the SVM, with NAP. The

[TABLE 1] PERFORMANCE OF GMM-UBM (GMM) COMPARED WITH GMM/SVM (GSL) SYSTEMS. PERFORMANCE OF GMM-UBM 2004 SYSTEM IS GIVEN FOR COMPARISON.

SYSTEM	EER (%)	minDCF (×100)
LIA04	9.36	4.21
GMM-UBM	8.47	3.94
GSL	6.88	3.37

This experiment is done on NIST-SRE 06, 1conv-1conv condition, English only, male set (694 target tests and 8299 nontarget tests).

[TABLE 2] PERFORMANCE OF GMM/UBM WITH FACTOR ANALYSIS (GMM-FA), GMM/SVM WITH FACTOR ANALYSIS (GSL-FA) AND GMM/SVM WITH NAP (GSL-NAP) SYSTEMS, COMPARED WITH THE UBM/GMM BASELINE.

SYSTEM	EER (%)	minDCF (×100)
BASELINE (GMM)	8.67	3.37
GMM-FA	4.55	1.59
GSL-FA	4.48	1.62
GSL-NAP	5.28	1.69

This experiment is done on NIST-SRE 06, 1conv-1conv, English only, male set (it includes 694 target tests and 829 nontarget tests).

common novelty of both approaches is to directly model the intersession mismatches, rather than to compensate for their effects. This mismatch modeling implies the use of large speech corpora, with, for example, several recordings of a given speaker using several different handsets. In the previous example, the focus is on the variability due to the handset. It is important to notice that, for both FA and NAP, the underlying problem is inside the supervector space, with a large dimension (more than 25,000 in the case of the systems presented in this article). Both approaches are implemented in ALIZE/SpkDet [15], [21].

Table 2 presents the results for the FA-based systems. Results are provided for the GMM-UBM (using the symmetrical variant of FA proposed in [21]) and for the GSL system using classical FA. The improvement using both methods is clearly emphasized with a minDCF and an EER reduced by a factor of about 2 compared with the baseline GMM-UBM reference system.

LONG DURATION TRAINING

The amount of speech available for training a speaker model is an important factor in terms of performance. To evaluate the impact of this factor on the performance of a modern speaker recognition system, we present two experiments: one gathered from the unsupervised training NIST-SRE condition and one from the long training condition.

[TABLE 3] PERFORMANCE DEPENDING ON THE TRAINING DURATION (ONE VERSUS THREE FILES) OF THE GSL-FA SYSTEM.

TRAINING DURATION	EER (%)	minDCF (×100)
ONE FILE (SHORT2-SHORT3)	2.96	1.35
THREE FILES (3CONV-SHORT3)	1.04	0.76

This experiment is done on NIST-SRE 08, short2-short3 condition versus 3conv-short3 condition, English only, male set (short2-short3 involves 439 target tests and 6,176 nontarget tests; 3conv-short3 involves 405 target tests and 4,905 nontarget tests).

[TABLE 4] PERFORMANCE WITH/WITHOUT UNSUPERVISED ADAPTATION AND WITH ORACLE ADAPTATIONS.

SYSTEM	EER (%)	minDCF (×100)
GMM-FA	4.55	1.59
GMM-FA-UNSUPERVISED	2.36	0.89
GMM-FA-ORACLE	1.62	0.50
GSL-FA	4.48	1.62
GSL-FA-UNSUPERVISED	2.27	0.81
GSL-FA-ORACLE	1.71	0.56

This experiment is done on NIST-SRE 06, 1conv-1conv, English only, male set (it includes 694 target tests and 829 nontarget tests).

LONG DURATION TRAINING

Table 3 presents the results in terms of EER and minDCF of two experiments using GSL-FA (GMM/SVM with FA) system, where only the training duration is different. Clearly, the use of three times more data for training a speaker model allows a drastic improvement in terms of EER (from 2.96% to 1.04%) as well as for the minDCF (from 1.35 to 0.76).

VERY LONG DURATION TRAINING

For several years, the NIST-SRE has included a long-duration training task. More recently, another task named the unsupervised adaptation mode has been proposed, which is a mix between traditional one conversation (1 conv) training and long training . In this task, the system is able to take advantage of the data gathered during use.

Several research teams have proposed various solutions for this unsupervised-training framework [22], [23]. The LIA has developed such an approach named continuous adaptation [24], where the test data are always integrated into the targeted speaker model with a weight based on a confidence measure. This approach was applied to both the GMM-UBM and the GSL systems, with or without the session variability techniques.

In this article, we focus on the oracle mode, where the system knows if a speech segment included in the training set of a given speaker belongs to this speaker. The results of the pure unsupervised mode are also provided for information. It is important to note that the relatively new unsupervised training systems obtain very impressive results on some corpora but show inconsistent results on other corpora. The use of the oracle (supervised) mode eliminates this inconsistency. In this experiment, the number of tests per model speaker and the number of target tests by speaker (useful tests for model adaptation) are variable. On average, there are 3.74 target tests per speaker model for the 264 target speakers, with 90 speaker models having zero target tests. For the remaining speakers, there is an average of 7.07 target tests.

Table 4 proposes a summary of the experimental results using unsupervised training. All the presented systems use FA. This table demonstrates that the amount of training data is a key factor in speaker recognition performance. With the oracle adaptation, the EER is divided by a factor between 2.6 and 2.8 and the minDCF by a factor between 2.9 and 3.2. Compared with the reference baseline system (2006 GMM-UBM system without FA), the gain is significantly larger: the EER decreases from 8.67% for the reference system to 1.62% for the GMM-UBM system with FA and the oracle adaptation mode.

LOOKING AT ERROR RATES AS A PROGRESS CRITERION

As we have shown in this article, current speaker recognition systems are able to deal with large and increasing amounts of training data, either to reduce the session mismatch problem or to increase the quality of the targeted speaker models. This has resulted in an impressive level of performance, with an EER of 2.3% for a task that is difficult. The potential for the presented approaches is large, as the EER is about 1.62% with the oracle. Moreover, the session mismatch techniques are recent and should be able to deal with larger corpora, increasing the number of variability factors they are able to model.

With the resulting error rates and additional weighted improvements, it seems legitimate to ask if speaker recognition can be viewed as a solved problem. Indeed, if increasing the amount of available data decreases error rates, is it useful to work on the speaker recognition engine, itself? The end part of this section tries to answer this question by looking at the different factors linked to the performance as measured during NIST evaluations.

THE MAIN OBJECTIVE OF THE NIST-SRE IS TO PROVIDE AN INTEGRATED FRAMEWORK FOR SCIENTIFICALLY EVALUATING THE APPROACHES AND SYSTEMS IN THE FIELD OF SPEAKER RECOGNITION: THE PARTICIPANTS WORK ON THE SAME CORPUS AND PROTOCOLS, THE SAME PERFORMANCE CRITERION, AND ARE TIME-SYNCHRONIZED BY THE CAMPAIGN SCHEDULE.

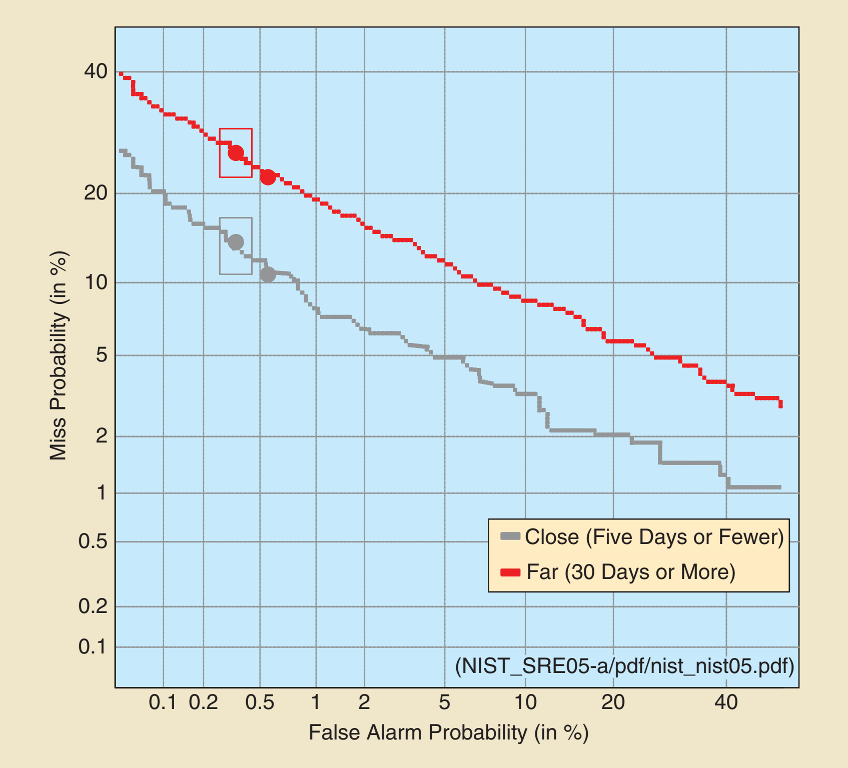
PERFORMANCE VARIABILITY

Multiple factors affect the performance of automatic speaker recognition systems, some depend on the speakers and others do not, while some factors can be difficult to isolate.

One factor hypothesized to affect performance is voice aging. Figure 1 (from NIST-SRE '05 [28], [33]) shows the impact of the elapsed time between recording the enrollment speech and the test speech. For a given, realistic threshold, the miss-probability error increases by a factor of two when the duration between enrollment and test exceeds one month; however, unfortunately, other factors were correlated with elapsed time, such as corpus collection bias (e.g., different proportions of non-English speakers in the two conditions). The hypothesis that factors other than voice aging are implied in the Figure 1 results is also supported by the fact that this very large aging loss was no longer noticed in SRE'06. Moreover, voice aging, and its effect on performance, is an ongoing research topic.

The duration and number of voice samples used in training are additional variables in system performance. In Table 5, we analyze the performance of the same LIA system, the GSL-FA, depending on the training conditions (short training based on one recording or long training based on three recordings) and the test subset (gender and language). All these results are extracted from the LIA NIST-SRE 2008 official participation. The variation factors have an important impact on the performance, raising EERs by up to a factor of 9 between two conditions. It is also interesting to observe that the differences are not consistent when several factors are moving. A part of this inconsistency is an artifact of the evaluation itself; there are fewer speakers and tests in the native speaker only condition than in the English one, for example. Moreover, these results clearly show that a unique error rate does not correctly describe the overall performance of a system.

Doddington et al. analyzed the impact of a set of variability factors on system performance, using NIST-SRE results [26]. In [27], the authors analyzed the results of the LIA GMM-FA



(NIST_SRE05-a/pdf/nist_nist05.pdf)

[FIG1] Effect of time between enrollment and test recordings, NIST-SRE '05.

[TABLE 5] RESULTS OF GSL-FA SYSTEM DEPENDING ON THE CORPUS SUBSET (EER% FOLLOWED BY MINDCF×100 IN PARENTHESES).

	SHORT2-SHORT3		3CONV-SHORT3	
	MALE	FEMALE	MALE	FEMALE
ALL LANGUAGE	5.95% (3.32)	8.54% (4.60)	3.67% (2.48)	6.58% (3.97)
ENGLISH RECORDS	2.96% (1.35)	3.54% (1.85)	1.04% (0.76)	2.05% (1.23)
ENGLISH RECORDS BY NATIVE SPEAKERS	0.89% (0.31)	2.14% (1.09)	2.10% (0.89)	3.42% (1.85)

This experiment is done on NIST-SRE'08, English only.

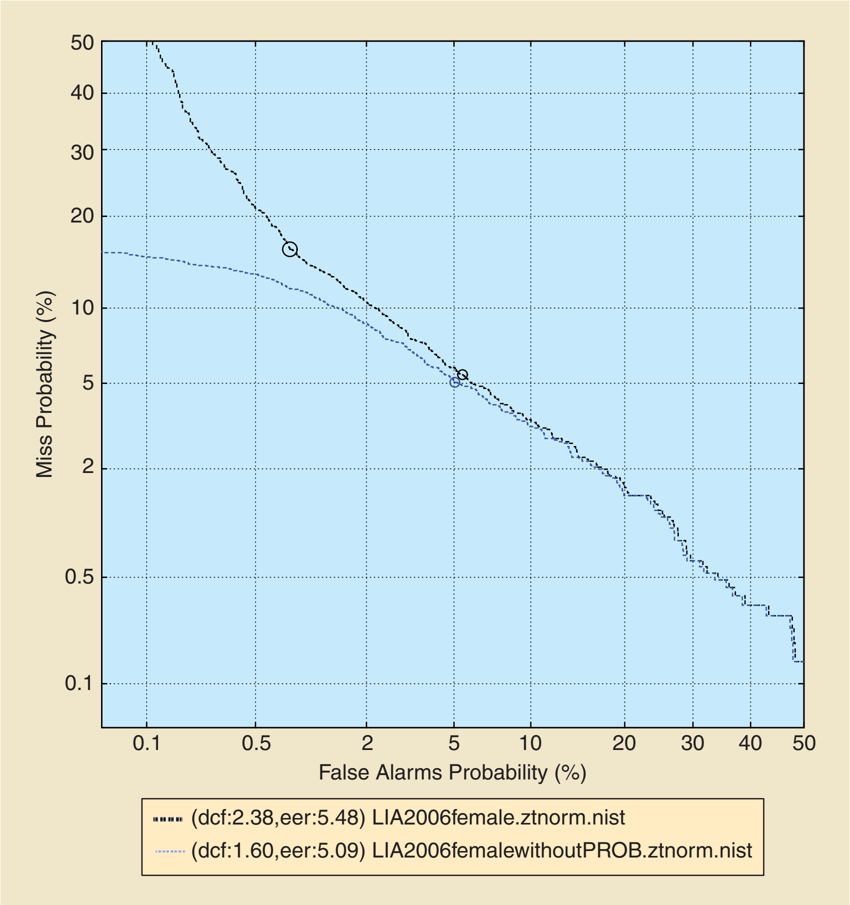
system, inside the NIST-SRE framework. They remark that a few impostor trials are responsible for about half of the system errors. Figure 2 shows two detection error trade-off (DET) curves: one computed using all the NIST protocol tests and the other one when less than 1% of the impostor trials are withdrawn (trials with the top scores are withdrawn). It is more interesting that the authors also show that the main part of this phenomenon is corrected when an inverse scoring is applied on the problematic tests (inverse scoring means that the speaker model is trained on the test file and scored against the enrollment file). This result demonstrates the suitability of the GMM-based approach with careful use of the training material.

[THE GAUSSIAN MIXTURE MODEL UNIVERSAL BACKGROUND MODEL APPROACH IS THE DOMINANT ONE IN TEXT-INDEPENDENT SPEAKER RECOGNITION.]

ERRORS IN CALIBRATION

Calibration [29], [1] is another significant issue that provokes caution when using automatic speaker recognition systems. Although a system may exhibit a low error rate, as indicated by its DET curve, it may be subject to different levels of variation in the actual score produced. The use of statistically significant similar tests can actually mask this issue. If the researcher has enough data from a single collection method, then the score of an automatic system can be observed and corrected for calibration errors. In some uses of automatic speaker recognition, this process may be appropriate, but in many forensic situations where the collection process is mismatched from enrollment to verification, good calibration may be difficult to achieve.

The speaker recognition community has taken significant steps to mitigate these issues. Compensation in classifier features (e.g., cepstral mean subtraction), model parameters (FA and NAP), and score normalization (T-norm, Z-norm) have all contributed to systems that have more predictable score distributions and, thus, can be calibrated. Significant work still remains, as evidenced by recent NIST evaluations. The cross-microphone task in NIST-SRE 2008 postevaluation showed that good error rates could be achieved, but calibration of systems varied dramatically as different cross-microphone types were examined. For example, for the same calibration technique and system of the Massachusetts Institute of Technology (MIT) Lincoln Laboratory (LL), in one cross-microphone case, a 5% relative error between the minimum and actual DCF was observed; in another case, a 160% change was observed. In both of these cases, the EER was below 2%. This observation demonstrates that calibration across significantly different conditions is still an area of research and affects the practical use of automatic speaker recognition systems.



[FIG2] Influence of a small subset of impostor trials on system performance (NIST-SRE'06, 1conv-1conv, all trial, female subset, GMM-FA system).

SPEAKER SPECIFIC INFORMATION: A DOUBT FACTOR

All the previous progress has been based on an underlying hypothesis: error rates are the criterion for evaluating both performance and progress in the speaker recognition field. In [25], the authors tried to show that this is not the only criterion, depending on the targeted scenario. They proposed to artificially transform the voices of the impostors to cheat a speaker recognition system (i.e., after the transformation, the system should recognize an impostor voice as coming from a targeted speaker). As the objective is to cheat a system and not a human expert, the only constraint at the human perception level is that the voice should remain natural. In this case, the transformation is done acoustically, frame by frame, and works only on the filter parameters of the classical source-filter model. The targeted speaker is known by the transformation system from an external (not included in the test protocol) extract of his or her voice and the targeted approach, GMM-UBM here. As the cepstral GMM-UBM system is used by all the methods presented in this article, it is reasonable to generalize the results of this experiment to all these presented systems. The experimental validation done in [25] is presented in Table 6.

> IN THE FORENSIC FIELD, THE ENVIRONMENT AND FACTORS AFFECTING PERFORMANCE CAN VARY TREMENDOUSLY, RELATIVE TO THE COMMERCIAL ARENA.

This transparent transformation technique introduces a significant factor into the speaker recognition system, as the false acceptance rate increases from less than 1% without the transformation to about 50% when the transformation is used.

These results give a new view to the impressive results observed recently in the speaker recognition field: have we really made drastic progress in speaker recognition? If it is possible to transform the voice of an impostor, with inaudible artificial modifications, and disrupt a speaker recognition system to such a great extent, is the information used by the system that user-specific?

This fact does not challenge the interest of the work done in the speaker recognition area. It is clear that significant progress has taken place in the last decade, e.g., in the session mismatch area, which was and remains a key challenge for speaker recognition.

COMMENTS

Nevertheless, this experiment, and others, shows that error rates might not be a sufficient criterion for evaluating both performance and progress in the speaker recognition field, even if it is necessary.

The amount of available speech material for both training and testing phases is important for the forensic context, where, quite often, only short pieces of speech are available. This constraint is known to have a large impact on speaker recognition performance. This aspect is not highlighted by NIST-SRE evaluations, even if some tasks on short durations are proposed. In [34], the authors investigate the effect of short durations on a GMM-UBM baseline system and on a GSL-NAP system, using the ALIZE/SpkDet software. They show that the EER of the GMM-UBM increases about a factor if 3 when only the duration of both training and testing speech excerpts is 10 s (the most difficult situation). The authors remark also that the new and very efficient session mismatch normalization techniques (FA and NAP) are very sensitive to the speech duration factor.

A DIFFERENT APPROACH TO SPEAKER RECOGNITION RESEARCH?

The aim of this article is to focus on the danger of using error rates as the only criterion for evaluating the state and the potential of speaker recognition research and technology. It is dangerous, both in terms of potential application and research orientation: there are aims other than the performance, as measured currently. This problem is more critical in the forensic field than in the commercial area. Commercial applications usually involve a clear application scenario where the environment and the variability factors are fairly well defined, or, at least, understood. In the forensic field, the environment and factors affecting performance can vary tremendously, relative to the commercial arena.

The evolution of speaker recognition, with a focus on error-rate reduction, progressively concentrates the research community on the engineering area, with less interest in the theoretical and analytical areas, involving phoneticians, for example. Nevertheless, it seems reasonable to develop automatic systems to aid in gaining a deeper understanding of the underlying phenomena. We propose the following solutions, which could extend knowledge in this field:

1) Analyze the performance on the phonetic information present in both the training and testing recordings. This corresponds to an analytical analysis of results, in terms of phonetic or linguistic content, to better understand which information is used by our systems. The use of artificial, well-controlled stimuli could be included in this study, and a comparison between machines and human perception seems very interesting in this case.

2) Work on more controlled data, possibly simulated data. It might be useful to start with a given natural or synthesized voice and to create various stimuli by working on each parameter, one by one (source parameters, filter parameters, prosody, vocalic triangle, formants, etc.). This work could include perceptual studies.

3) Performance evaluation should integrate more variability factors, more heterogeneous factors, and more unknown factors to allow a better generalization of the results. Doing that with real-world recorded data is certainly very costly, but using voice transformation and voice synthesis techniques open more practical solutions. Even if this solution is of great interest, it remains true that the only scientifically strong solution is to increase the size of the evaluation corpora and protocols.

4) The evaluation is currently based on recordings involving hundred of speakers and on thousands of speaker

[TABLE 6] EFFECT OF ARTIFACT-FREE IMPOSTOR VOICE TRANSFORMATION.

	BASELINE SYSTEM	BASELINE SYSTEM + ARTIFICIAL IMPOSTOR VOICE TRANSFORMATION
EER (%)	8.54	35.41
minDCF (×100)	3.58	9.41
FALSE ACCEPTANCE (%)	0.88	49.72
FALSE REJECT (%)	27.45	27.45

This experiment is done on NIST-SRE'06, 1conv-1conv, all trials, male only.

recognition tests. Changing the size factor, to have thousands of speakers, hundred thousands of tests, and hundreds of mixed conditions, is the best way to achieve a strong, unquestionable evaluation of performance and progress.

5) Although current systems achieve good results, they also show some surprises, like some impostor tests that obtain very high scores (higher than the mean of target speaker tests). The number of such tests is very small, about one hundred for a NIST evaluation, but this number is relative given the small size of the evaluation. Giving higher importance to isolated and unusual results could constitute an easy and interesting way to change the focus of speaker recognition research.

6) As mentioned in the previous point, the current systems work very well, in general, but not in all the operational environments. Our main paradigm is based on statistical modeling and analysis, so it is usually difficult to detect these problems because they are, by nature, rare. An alternative is to work more on the nature of the information present in different recordings, to predict if one recording corresponds or not to the underlying hypothesis linked to a specific speaker recognition approach.

[MULTIPLE FACTORS AFFECT THE PERFORMANCE OF AUTOMATIC SPEAKER RECOGNITION SYSTEMS, SOME DEPEND ON THE SPEAKERS AND OTHERS DO NOT, WHILE SOME FACTORS CAN BE DIFFICULT TO ISOLATE.]

CONCLUSIONS

Looking at the different points highlighted in this article, we affirm that forensic applications of speaker recognition should still be taken under a necessary need for caution. Disseminating this message remains one of the most important responsibilities of speaker recognition researchers.

ACKNOWLEDGMENT

This work is sponsored, in part, by the Department of Homeland Security under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

AUTHORS

Joseph P. Campbell (j.campbell@ieee.org) received a B.S.E.E. degree from Rensselaer Polytechnic Institute in 1979, an M.S.E.E. degree from Johns Hopkins University (JHU) in 1986, and a Ph.D. degree from Oklahoma State University in 1992. He is currently senior staff at MIT-LL in the Information Systems Technology Group, where he conducts speech-processing research and specializes in speaker recognition and biometrics. Before joining MIT-LL, he served 22 years at the National Security Agency. He chaired the Biometric Consortium from 1994 to 1998, taught JHU's graduate course Speech Processing from 1991 to 2001, and coedited *DSP Journal* from 1998 to 2005. He is a cochair of the International Speech Communication Association's Speaker and Language Characterization Special Interest Group as well as a member of the National Academy of Sciences' Whither Biometrics? Committee. He serves on the IEEE Kilby Medal Committee and is vice president of Technical Activities of the IEEE Biometrics Council. Dr. Campbell is a member of Sigma Xi, the International Speech Communication Association, the Boston Audio Society, and the Acoustical Society of America. He is a Fellow of the IEEE.

Wade Shen (swade@ll.mit.edu) is currently with MIT Lincoln Laboratory. He received his master's degree in computer science from the University of Maryland, College Park in 1997, and his bachelor's degree in electrical engineering and computer science from the University of California, Berkeley in 1994. His current areas of research involve machine translation and machine translation evaluation, speech, speaker, and language recognition for small-scale and embedded applications, named-entity extraction, and prosodic modeling. Prior to joining Lincoln Laboratory in 2003, Shen helped found and served as Chief Technology Officer for Vocentric Corporation, a company specializing in speech technologies for small devices.

William M. Campbell (wcampbell@ll.mit.edu) is a technical staff member in the Information Systems Technology group at Massachusetts Institute of Technology (MIT) Lincoln Laboratory. He received his Ph.D. in applied mathematics from Cornell University in 1995. Prior to joining MIT Lincoln Laboratory, he worked at Motorola on biometrics, speech interfaces, wearable computing, and digital communications. His current research interests include machine learning, speech processing, and social network analysis.

Reva Schwartz (reva.schwartz@usss.dhs.gov) is currently a national expert and forensic examiner at the Forensic Services Division of the United States Secret Service, where she serves as the expert in the conduct of research and analysis in speech and signal processing, forensic speaker recognition, and the enhancement of audio recordings, and other forensic evidence. She also serves as project manager for federally funded research programs and provides expert advice and guidance within the agency and to other federal, state, local, and international law enforcement agencies concerning speech processing and audio enhancement as investigative and intelligence techniques. She is a member of the American Academy of Forensic Sciences, International Association for Forensic Phonetics and Acoustics, Acoustical Society of America, and the Forensic Speech, Audio & Authentication Working Group of the European Network of Forensic Science Institutes.

Jean-François Bonastre (jean-francois.bonastre@univ-avignon.fr) obtained his Ph.D. degree in 1994, in Avignon, France, in automatic speaker identification using phonetic-based knowledge. He then joined the LIA (University Avignon, France) as an associate professor and became full professor in 2008. As a member of the Natural Language Processing Group, he developed his research in speaker characterization and recognition using phonetic, statistic and prosodic information, while teaching and lecturing on various subjects covering computer science, speech processing, audio signal classification and indexing, and biometry. From 2001 to 2004, he was the chairman of AFCP, the French-Speaking Speech Communication Association (currently a regional branch of ISCA). He was also the chair of the ISCA Speech and Language Characterization SIG for two years and he joined the board of ISCA in 2005. He has been vice president of ISCA since 2007. He is a member of the IEEE Speech and Language Technical Committee. He was also a member of the technical committee of several conferences (and area chair for Interspeech 2008). He is a Senior Member of the IEEE.

Driss Matrouf (driss.matrouf@univ-avignon.fr) obtained his Ph.D. degree in 1997 in noisy speech recognition, from Paris IX University. He then joined LIA (University Avignon, France), as an associate professor. His research interests include speech recognition, language recognition, and speaker recognition. His research interests are session and channel compensation for speech and speaker recognition. Parallel with his research activities, he teaches in the fields covering computer science, speech coding, and information theory. He is also responsible for the University of Avignon computer science bachelor program. He is a Member of the IEEE.

REFERENCES

[1] W. Campbell, K. Brady, J. Campbell, D. Reynolds, and R. Granville, "Understanding scores in forensic speaker recognition," in *Proc. IEEE Odyssey, ISCA Speaker Recognition Workshop*, June 28–30, 2006, pp. 1–8.

[2] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Proc. Eurospeech, ISCA*, Geneva, Switzerland, 1–4 Sept., 2003, pp. 33–36.

[3] R. H. Bolt, F. S. Cooper, D. M. Green, S. L. Hamlet, J. G. McKnight, J. M. Pickett, O. Tosi, B. D. Underwood, and D. L. Hogan, *On the Theory and Practice of Voice Identification*. Washington, D.C.: National Research Council, National Academy of Sciences, 1979.

[4] R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes," *J. Acoust. Soc. Amer.*, vol. 47, no. 2, pp. 597–612, 1970.

[5] J. F. Nolan, *The Phonetic Bases of Speaker Recognition*. Cambridge, MA: Cambridge Univ. Press, 1983.

[6] L. J. Boë, "Forensic voice identification in France," *Speech Commun.*, vol. 31, no. 2–3, pp. 205–224, June 2000.

[7] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[8] D. A. Reynolds, W. D. Andrews, J. P. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusácek, J. S. Abramson, R. Mihaescu, J. J. Godfrey, D. A. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, 2003, pp. 784–787.

[9] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, 2002, pp. 300–304.

[10] A. Martin and M. Przybocki. The NIST speaker recognition evaluation series, National Institute of Standards and Technology's Web site [Online]. Available: http://www.nist.gov/speech/tests/sre

[11] P. Kenny and P. Demouchel, "Eigenvoices modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, 2005.

[12] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2006, pp. I-97–I-100.

[13] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovksa-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 4, pp. 430–451, 2004.

[14] 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. IEEE Odyssey, ISCA Speaker Recognition Workshop*, 2008.

[15] B. G. B. Fauve , D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1960–1968, Sept. 2007.

[16] W. M. Campbell, D. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, May 2006, pp. 308–311.

[17] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2003, pp. II-53–II-56.

[18] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. IEEE Odyssey, ISCA Speaker Recognition Workshop*, 2001, pp. 213–218.

[19] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Dig. Signal Process.*, vol. 10, no. 1–3, pp. 42–54, 2000.

[20] A. Solomonoff, M. W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, pp. 629–632.

[21] D. Matrouf,N, Schefferm B. Fauve, and J.-F. Gauvain, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech*, 2007.

[22] C. Barras, S. Meignier, and J.-L. Gauvain, "Unsupervised online adaptation for speaker verification over the telephone," in *Proc. IEEE Odyssey, ISCA Speaker Recognition Workshop*, 2004, pp. 157–160.

[23] D. Van Leeuwen, "Speaker adaptation in the NIST speaker recognition evaluation 2004," in *Proc. Interspeech*, 2005, pp. 1981–1984.

[24] A. Preti, J.-F. Bonastre, F. Capman, and B. Ravera, "Confidence measure based unsupervised target model adaptation for speaker verification," in *Proc. Interspeech*, 2007.

[25] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, 2007, pp. 2053–2056.

[26] G. R. Doddington, W. Liggett, A. Martin, M. Przybocki, D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1998, pp. 1351–1354.

[27] S. E. Mezaache, J.-F. Bonastre, and D. Matrouf, "Analysis of impostor tests with high scores in NIST-SRE context," in *Proc. Interspeech*, 2008.

[28] NIST. 2005 speaker recognition evaluation [Online]. Available: http://www.nist.gov/speech/tests/sre/

[29] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 230–275, 2006. [Online]. Available: http://www.informatik.uni-trier.de/%7Eley/db/journals/csl/csl20.html#BrummerP06

[30] R. Schwartz, "Voiceprints in the United States—Why they won't go away," in *Proc. Int. Association for Forensic Phonetics and Acoustics*, Göteborg, Sweden, 2006.

[31] F. Nolan and T. Oh, "Identical twins, different voices," *Forensic Linguistics*, vol. 3, no. 1, pp. 39–49, 1996.

[32] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluation chronicles, Part 2," in *Proc. IEEE Odyssey, ISCA Speaker Recognition Workshop*, 28–30 June, 2006, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4013537

[33] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora—2004, 2005, 2006," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 1951–1959, Sept. 2007. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4291612

[34] B. Fauve, N. Evans, N. Pearson, J. F. Bonastre, and J. S. D. Mason, "Influence of task duration in text-independent speaker verification," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 794–797.

[35] P. Rose, *Forensic Speaker Identification*. London: Taylor & Francis, 2002.

[36] A. F. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection performance," in *Proc. Eurospeech, Rhodes*, Greece, 1997, pp. 1895–1898 [Online]. Available: http://www.nist.gov/speech/publications/