



Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014)

Joaquin Gonzalez-Rodriguez

ATVS-Biometric Recognition Group, Universidad Autónoma de Madrid

e-mail: joaquin.gonzalez@uam.es

Citation / Cómo citar este artículo: Gonzalez-Rodriguez, J. (2014). Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014). *Loquens*, 1(1), e007. doi: <http://dx.doi.org/10.3989/loquens.2014.007>

ABSTRACT: Automatic Speaker Recognition systems show interesting properties, such as speed of processing or repeatability of results, in contrast to speaker recognition by humans. But they will be usable just if they are reliable. Testability, or the ability to extensively evaluate the goodness of the speaker detector decisions, becomes then critical. In the last 20 years, the US National Institute of Standards and Technology (NIST) has organized, providing the proper speech data and evaluation protocols, a series of text-independent Speaker Recognition Evaluations (SRE). Those evaluations have become not just a periodical benchmark test, but also a meeting point of a collaborative community of scientists that have been deeply involved in the cycle of evaluations, allowing tremendous progress in a specially complex task where the speaker information is spread across different information levels (acoustic, prosodic, linguistic...) and is strongly affected by speaker intrinsic and extrinsic variability factors. In this paper, we outline how the evaluations progressively challenged the technology including new speaking conditions and sources of variability, and how the scientific community gave answers to those demands. Finally, NIST SREs will be shown to be not free of inconveniences, and future challenges to speaker recognition assessment will also be discussed.

KEYWORDS: automatic speaker recognition; discrimination and calibration; assessment; benchmark

RESUMEN: *Evaluando los sistemas automáticos de reconocimiento de locutor: Panorama de las evaluaciones NIST de reconocimiento de locutor (1996-2014).*- Los sistemas automáticos de reconocimiento de locutor son críticos para la organización, etiquetado, gestión y toma de decisiones sobre grandes bases de datos de voces de diferentes locutores. Con el fin de procesar eficientemente tales cantidades de información de voz, necesitamos sistemas muy rápidos y, al no estar libre de errores, lo suficientemente fiables. Los sistemas actuales son órdenes de magnitud más rápidos que tiempo real, permitiendo tomar decisiones automáticas instantáneas sobre enormes cantidades de conversaciones. Pero tal vez la característica más interesante de un sistema automático es la posibilidad de ser analizado en detalle, ya que su rendimiento y fiabilidad puede ser evaluada de manera ciega sobre cantidades enormes de datos en una gran diversidad de condiciones. En los últimos 20 años, el Instituto Nacional de Estándares y Tecnología (NIST) de EE. UU. ha organizado, proporcionando los datos de voz y protocolos de evaluación adecuada, una serie de evaluaciones de reconocimiento de locutor independiente del texto. Esas evaluaciones se han convertido no sólo en una prueba comparativa periódica, sino también en punto de encuentro de una comunidad colaborativa de científicos que han estado profundamente involucrados en el ciclo de evaluaciones, lo que ha permitido un enorme progreso en una tarea especialmente compleja en la que la información individualizadora del locutor se encuentra dispersa en diferentes niveles de información (acústica, prosódica, lingüística...) y está fuertemente afectada por factores de variabilidad intrínsecos y extrínsecos al locutor. En este artículo se describe cómo las evaluaciones desafiaron progresivamente la tecnología existente, incluyendo nuevas condiciones de habla y fuentes de variabilidad, y cómo la comunidad científica fue dando respuesta a dichos retos. Sin embargo, estas evaluaciones NIST no están libres de inconvenientes, por lo que también se discutirán los retos futuros para la evaluación de tecnologías de locutor.

PALABRAS CLAVE: reconocimiento automático de locutor; calibración; validación; prueba de referencia

1. INTRODUCTION

The massive presence and exponential growth of multimedia data, with audio sources varying from call centers, mobile phones and broadcast data (radio, TV, podcasts...) to individuals producing voice or video messages with speakers and conversations of all types in an unconceivable range of situations and speaking conditions, makes speaker recognition applications critical for the organization, labeling, management and decision making over this *big audio data*. In order to efficiently process such amounts of speech information, we need extremely fast and reliable enough (being not error free) automatic speaker recognition systems. Current systems are orders of magnitude faster than real time, allowing instantaneous automatic decisions (or informed opinions) on huge amounts of conversations. Moreover, being built with well-known signal processing and pattern recognition algorithms, they are *transparent*, avoiding subjective components in the decision process (in contrast to human speaker recognition) and allowing public scrutiny of every module of the system, and *testable*, as their performance and claimed reliability can be blindly evaluated on massive amounts of known data where ground truth (the speaker identity) is available in a great diversity of conditions (Gonzalez-Rodriguez, Rose, Ramos, Toledano, & Ortega-Garcia, 2007).

However, evaluating a speaker recognition system is not an easy task. The speaker individualizing information is spread across different information levels, and each of them is affected in different ways by speaker intrinsic and extrinsic sources of variability, such as the elapsed time between recordings under comparison, differences in acquisition and transmission devices (microphones, telephones, GSM/IP coding...), noise and reverberation, emotional state of the speaker, type of conversation, permanent and transient health conditions, etc. When two given recordings are to be compared, all those factors are empirically combined in a specific and difficult to emulate way, so controlling and disentangling them will be critical for the proper evaluation of systems. Fortunately, the US NIST (National Institute of Standards and Technology) series of Speaker Recognition Evaluations (SRE) have provided for almost two decades a challenging and collaborative environment that has allowed text-independent speaker recognition to make remarkable progress. For every new evaluation, and according to the results and expectations from the previous one, NIST balanced the suggestions from participants and the priorities of their sponsors to compound a new evaluation, where new speech corpora were designed and collected under new conditions, development and test data was prepared and distributed, and final submissions from participants were analyzed, to be finally compared and discussed in a public workshop open to participants in the evaluation. This cycle of innovation adds a tremendous value to participants, whose technology is challenged in every new evaluation, demanding enormous progress in order to cope with new demands, and final solutions being publicly discussed and scruti-

nized with other participants, creating an environment for mutual development and enrichment.

This paper, without going into deep technical details but with proper selected references for the interested reader, pretends to be an easy-to-read guided tour over the different technologies and tasks that have been evolving jointly in the last two decades in the challenging world of text-independent speaker recognition. Trying to do so, the paper is organized as follows. After this introduction, section 2 depicts the different configurations and options available to build and deploy an automatic speaker recognition system. Section 3 deals with the evaluation of the goodness of a given speaker recognizer, from the design and preparation of reference data to cost functions and application-independent assessment of speaker detectors. The next four sections perform an historic overview of technologies and evaluations, from early short-term spectral systems and later higher-level systems to factor analysis and state-of-the-art i-vector systems, with the corresponding evaluations that challenged each of these systems. Finally, we extract some conclusions to summarize the paper and discuss relevant issues.

2. FLAVOURS IN AUTOMATIC SPEAKER RECOGNITION

This section gives an introductory outlook of the different architectures and speaker information extraction components that can be used to build an automatic system. Interested readers will find in each section a selection of references that allows going into further details in the different aspects addressed in the paper.

2.1. Text-dependent and text-independent systems

Automatic systems can be divided into two big groups, depending on the level of dependence with the pronounced message in the speech under comparison. *Text-independent* systems are focused on the different sounds produced by the speaker, independently from the language being spoken allowing cross-language speaker comparisons, and the message being pronounced, allowing comparisons of totally different utterances with different messages, speaking contexts and conditions, etc. Those systems allow comparisons of any two given utterances, but in order to be reliable they require significant amounts of spoken material (usually more than 30 seconds per utterance) and roughly similar acoustic and speaking conditions. The higher the mismatch between the conditions from one recording to the other (in manner of speaking, recording channel and acoustic noise, time lapse between recordings, etc.), the greater the degradation in performance. Those systems, extensively reviewed in Kinnunen and Li (2010), will be the subject of analysis in detail in this paper.

On the other side, *text-dependent* systems require the two utterances under comparison to pronounce ex-

actly the same words or phrases. At the expense of this constraint, by controlling the linguistic variability (same set of prototype sounds to be compared and in the same sequential order) they can obtain excellent levels of reliability with very short phrases/passwords from cooperative speakers. Text-dependent systems are especially suited to biometric access control applications, such as remote phone banking or customer phone platforms, where a spoken username and/or password are required.

One of the vulnerabilities of those systems is fraudulent recording of the spoken password, which can be avoided by random prompts to the speaker. In that case, the system must be flexible enough to build in real-time composite phrase models of the speaker from previously trained basic linguistic units (digits, phones, diphones...). In this way, the speaker is verified with a new phrase he has never recorded before every time he or she accesses the system. Details on performance levels with different architectures and design options can be found in the excellent review chapter on text-dependent systems in Hébert (2008).

Recently, a renewed interest in those systems has been observed, as shown by the MOBIO evaluation (Khouri et al., 2013) for voice access control in mobile environments whose results were presented at the International Conference on Biometrics (ICB) 2013 in Madrid, and in an INTERSPEECH 2014 special session entitled “Text-dependent speaker verification with short utterances” which will be focused on “robustness with respect to duration and modeling of lexical information” (Larcher, Aronowitz, Lee, & Kenny, 2014). Interestingly, the organizers have made publicly available the RSR2015 database, including 150 hours of data recorded from 300 speakers in mobile environments, which allow text-dependent system design and evaluation in different configurations, and comparison of results with other systems using the same database (Larcher, Lee, Ma, & Li, 2014).

2.2. Multi-level extraction of the individualizing information

From the particular realization of speech sounds to the production of spoken language (selection of words, phrase formulation, etc.), going through particular prosodic contours or voice qualities, every speech act embeds information from the speaker at multiple levels. While most automatic systems rely on short-term *cepstral*-like features—MFCC (Davis & Mermelstein, 1980), RASTA-PLP (Hermansky & Morgan, 1994), etc.—which will be the main information extraction techniques throughout this paper, there is a wide corpus of research in non-cepstral features.

Voice-source features representing the glottal information in the speech signal have also been extracted with success and used to improve current performance of cepstral systems (Plumpe, Quatieri, & Reynolds, 1999). However, the difficulty in correctly extracting those glottal and source features has resulted in limited

improvements of performance when combined when cepstral-like features. Fortunately, a recent software repository known as COVAREP (Degottex, Kane, Drugman, Raitio, & Scherer, 2014) provides state-of-the-art open-source glottal and voice source extraction tools which will surely help to improve their contribution to global performance.

In order to capture the characteristic *coarticulation* of the speakers in 100 to 500 milliseconds window lengths, different *spectro-temporal* features have been explored with success. Among them we can highlight the representation of the spectral variations as a function of time as frequency filtered spectral energies (Hernando & Nadeu, 1998) or frequency modulation features (Thiruvanan, Ambikairajah, & Epps, 2008). Recently, the trajectories of formant frequencies and bandwidths within specific linguistic units (phones, diphones, triphones, syllables and words) have been exploited obtaining a compact fixed-size representation of the formant dynamics per linguistic unit, obtaining good speaker recognition results just from formants in the NIST 06 framework (Gonzalez-Rodriguez, 2011).

Prosodic information contains characteristic speaker features embedded in pitch and energy contours, which can be tokenized (a discretization of the joint pitch-energy tendencies) and modeled through n-gram counts (Adami, Mihaescu, Reynolds, & Godfrey, 2003). With the help of an automatic speech recognition system, which provides precise phone boundaries in the input utterance, syllable, phone and state-in-phone durations can also be modeled (Shriberg, 2007). Recently, Kockmann, Ferrer, Burget, Shriberg, and Černocký (2011) have integrated state-of-the-art i-vectors (see Section 6.2) with prosodic information with important improvements over spectral-only systems in very complex tasks.

A statistical approach can also be used to extract *idiolectal* features from the speaker, looking for the frequency of use of bi-grams and tri-grams of phone, syllables or words (Doddington, 2001). This information, which combines extremely well with short-term cepstral approaches, becomes really useful only when large amounts of voice from different conversations of the speaker are available, as e.g., eight or sixteen five-minute two-sided conversations in NIST 04, which gives an average of 20 or 40 minutes per speaker. This minimum duration limit can be a severe drawback for some applications, but there are situations when those amounts of speech or much more are available, as for instance frequent users of customer call centers, or weeks or months of wiretapping, where hours of conversations are available in plenty of criminal investigations.

We have to highlight that some of the non-cepstral methods described make use of phone, syllable or word transcriptions automatically provided by *Automatic Speech Recognition* (ASR) systems. Those time-aligned labels are then used for phone, diphone, triphone, syllable or word selection and/or conditioning of specific speech segments including relevant information to the system in use. During almost a decade,

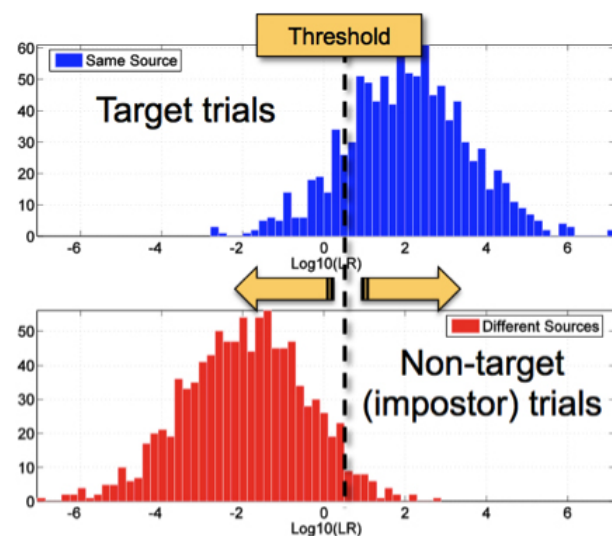
NIST have provided participants with errorful (15-30% of word error rate) word transcriptions of conversations, but suddenly interrupted this policy for the 2012 evaluation.

3. ASSESSMENT OF SPEAKER RECOGNITION SYSTEMS

One of the main advantages of automatic systems is that they can be extensively and repeatedly tested to assess their performance in a variety of evaluation conditions, allowing objective comparison of systems in exactly the same task, or observing the performance degradation of a given system in progressively challenging conditions. In order to assess a system, we need a database of voice recordings from known speakers, where the ground truth about the identity of the speaker in every utterance is known in order to compare it with the blind decisions of the system.

At this point, we need some definitions. A *test trial* will consist in determining if a given speaker (typically the speaker in a control recording) is actually speaking in the test recording. We will talk about *target* trials when the target (known) speaker is actually speaking in the test recording, and *non-target* (or *impostor*) trials in the opposite condition (the speakers in the train and test recordings are different). Automatic systems, given a test trial of unknown-solution, provide a score; the higher the score the greater the confidence in being same-speaker recordings. In an ideal system, the distribution of target-trial scores should be clearly separated and valued higher than that of non-target trials, allowing perfect discrimination setting a threshold between the two distributions of scores. However, as shown in Figure 1, target and non-target score distributions usually overlap each other partially.

Figure 1: Overlapping histograms of target and non-target score distributions.

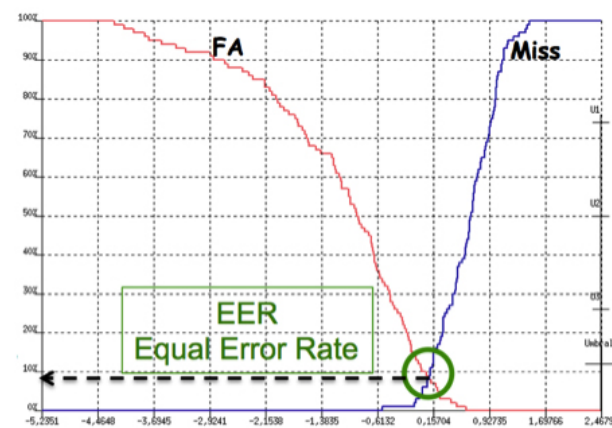


A detection threshold is needed if we want the system to provide hard acceptance (score higher than the threshold) or rejection (score lower than the threshold) decisions for every trial. Two types of errors can be committed by the system: *false alarm* (or *false acceptance*) errors, if the score is higher than the threshold in a non-target trial, and *miss detections* (also called false rejections) if the score is lower than the threshold in a target trial. Then, as the score distributions overlap, whatever the detection threshold we set a percentage of false acceptances and missed detections will occur: the higher the threshold, the lower the false alarms and the higher the miss detections. This means that, for any given system, lots of operation points are possible, with different values of compromise between both types of error.

3.1. ROC and DET curves

In Figure 2, both false acceptance and miss detection errors are shown as a function of a sweeping threshold. The error rate obtained where both false acceptance and miss detection curves cross each other is known as the Equal Error Rate (EER), and it is commonly used as a single number characterizing system performance: the lower the EER the better the separation between both target and non-target score distributions.

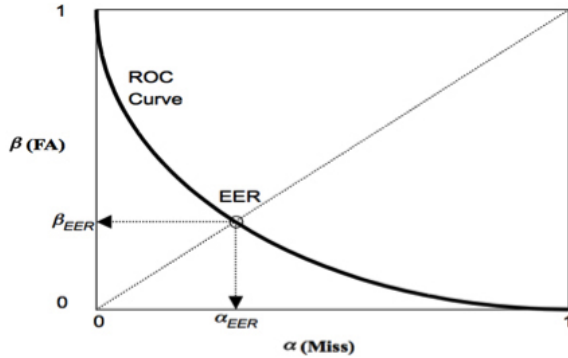
Figure 2: Percentage (%) of false alarms and miss detections as function of a sweeping threshold (x-axis), and Equal Error Rate (EER).



However, multiple error curves with different shapes, possibly overlapping and crossing each other, can result in the same EER, making it difficult to compare different systems. An alternative single curve representing all possible operating points of the system is usually preferred, as shown in Figure 3, known as Receiver Operating Characteristic (ROC) curve, where all the pairs of errors (false alarm, miss) for a sweeping threshold are represented as points in the false acceptance versus missed detection plane, and connected with a single line. The advantage of a ROC plot is that it represents all the

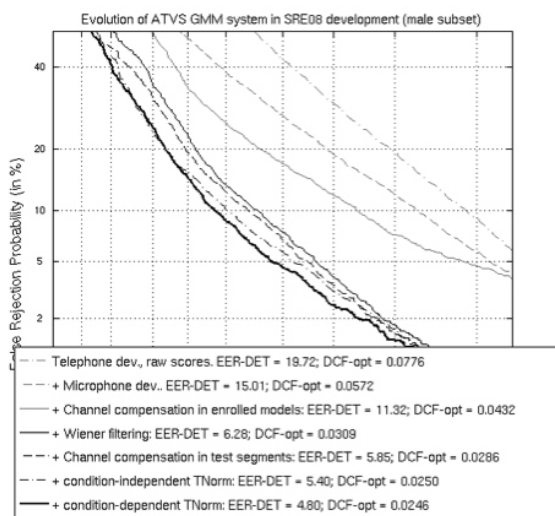
operating points of a system in a single curve, allowing easy comparison of systems (the closer to the origin of coordinates, the better). Moreover, the commonly used EER point is easily extracted where the ROC curve crosses the diagonal straight line where false acceptance errors equal the missed detection errors.

Figure 3: Receiver Operating Characteristic (ROC) curve summarizing false alarm and miss detection curves. EER (%) is easily obtained where the ROC curve crosses the FA=miss diagonal.



However, when different systems (or the same system in different evaluation conditions) with low EER are represented, the graphical resolution is very low as all the useful information is concentrated in the lower corner close to the origin of coordinates. The Detection Error Trade-off (DET) plot is suggested to overcome this problem, modifying the axis in order for Gaussian distribution of scores to result in straight lines in the DET plot, allowing for easy comparison and detailed analysis of different possible operation points of the system, as shown in Figure 4.

Figure 4: Sample DET curves from the development phase of ATVS-UAM systems for SRE 2008. The telephone-based SRE 2006 GMM-UBM system is progressively adapted to the cross-channel mic and tel condition in 2008.



3.2. Cost function

In Section 3.1 we have shown how to evaluate the goodness of the discrimination abilities of a system. However, equally interesting is to know about the goodness of the decisions (acceptance/rejections) of a system for a given application of interest. Applications are characterized by two factors: the relative cost of a false alarm versus a miss detection, and the *a priori* probability of observing target and non-target speakers. We can illustrate those two factors with opposite applications where we could use exactly the same speaker recognition system with differing detection thresholds. Firstly, imagine a non-critical user-friendly application, of little interest to attackers or impostors. Here, the *a priori* probability of being a true user is high ($P_{\text{target}} \geq P_{\text{non-target}}$), and system designer will look not to disturb them in excess, so $C_{\text{miss}} \geq C_{\text{FA}}$. If in contrast we have a high-security application, costs will verify $C_{\text{FA}} \gg C_{\text{miss}}$. If we look for a given speaker in big data repositories, where the probability of finding the target speaker is very low, *a priori* probabilities will verify $P_{\text{target}} \ll P_{\text{non-target}}$.

For a given discrimination capability, depending on the selected threshold the system can observe different probabilities of false alarm, $P_{\text{FA}|\text{non-target}}$, and miss detection, $P_{\text{miss}|\text{target}}$. NIST have used for their Speaker Recognition Evaluations (SRE) a composite cost function in order to rank participant systems where, given the application-dependent parameters C_{miss} , C_{FA} and P_{target} ($P_{\text{non-target}} = 1 - P_{\text{target}}$), and the system-dependent probabilities $P_{\text{FA}|\text{non-target}}$ and $P_{\text{miss}|\text{target}}$, an objective cost function C_{DET} can be computed:

$$[1] \quad C_{\text{DET}} = C_{\text{miss}} \cdot \frac{P_{\text{target}}}{P_{\text{FA}|\text{non-target}}} \cdot P_{\text{miss}|\text{target}} + C_{\text{FA}} \cdot P_{\text{non-target}}$$

It is important to note that C_{DET} can only be computed *a posteriori* once the correct labels are known, that is, after the evaluation session. This means that participant systems have to select in advance their decision threshold for the given application parameters, and once results are submitted, NIST tells participants their actual C_{DET} . But once the solution labels are known, it is also possible to compute which was the optimal threshold of the system for the given discrimination, which provides the $\min C_{\text{DET}}$ value. Systems with close values of C_{DET} and $\min C_{\text{DET}}$ are well calibrated, while the bigger the difference, the bigger the calibration loss. We have to highlight that it is possible to have a very good discriminant system with a very bad performance at the evaluation because of a bad threshold selection. Therefore, threshold selection is maybe the most critical part of a NIST submission, as it has the biggest direct impact in the system's associated cost.

However, submitting system scores in the form of calibrated likelihood ratios is strongly recommended. Doing so, there is no longer a need to set any detection threshold, as the detection threshold is directly obtained

from these application-parameters. The system will be then application independent, meaning it is valid for any given application (access-control, customer-oriented, forensic...) whatever the application-parameters (a priori probabilities of targets, and relative costs) (Brümmer, 2010; Ramos, Gonzalez-Rodriguez, Zadora, & Aitken, 2013; van Leeuwen & Brümmer, 2007). Additionally, this is extremely useful for applications such as forensics (Gonzalez-Rodriguez et al., 2007; Rose, 2002) where priors and costs are unknown (case- and legal-system dependent) and the reported likelihood ratios must be valid for whatever priors and costs are given (if any).

4. THE EARLY NIST SRES (1996-2001): ACOUSTIC-SPECTRAL SHORT-TERM SYSTEMS

Text-independent speaker recognition technology in the 90's was dominated by the GMM-UBM approach (Reynolds, Quatieri, & Dunn, 2000). Under the assumption of frame independence, the short-term cepstral vectors representing the spectral content every (typically) 10 milliseconds are modeled through a generative model consisting in a weighted combination of multivariate Gaussian mixtures. This Gaussian Mixture Model (GMM) is expected to represent the observed features from the speaker in a cumulative way, which are taken as observations from this underlying model. GMMs, with high number of gaussians, were originally trained as maximum likelihood estimates from the training data, producing a large dependence for optimal systems on the duration of the available training material, as the number of free parameters to be estimated from data grows with higher number of gaussians. It was soon observed that a GMM representing all the available feature vectors from a given reference population, known as Universal Background Model (UBM) was very useful to normalize the score from the GMM, adding substantial robustness to duration variability. Finally, Maximum a Posteriori (MAP) adaptation of the speaker GMM from the UBM, where only the means of the gaussians are adapted to the speaker data, became the state-of-the-art at the time, being also the reference system to be compared with for almost two decades.

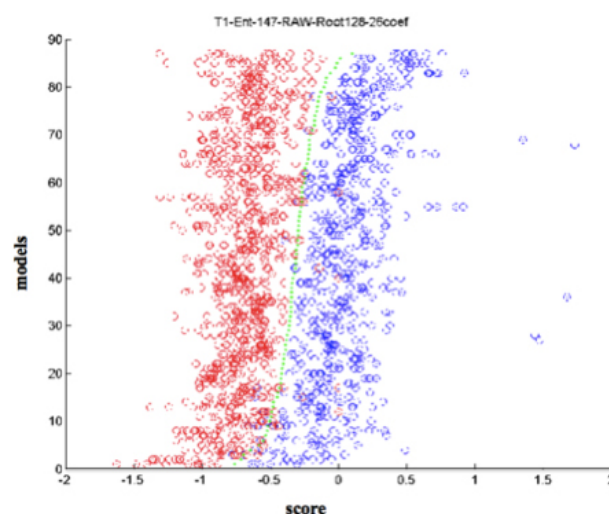
But those GMM-UBMs were severely affected by different sources of variability, such as available durations for training and test, type of telephone handset (carbon vs. electret), or land-line versus cellular phone calls. Those early evaluations were then designed to explore the limits and assess the performance of those systems in a variety of conditions to be sketched here.

4.1. Task definitions, corpora and evaluation conditions

Different tasks were explored in those early evaluations, but *one-speaker detection*, the task of determining

if a given speaker is actually speaking in a given conversation side, has always been the main subject of evaluation. For every trial, participants are required to submit both a score indicating the confidence in being the same speakers, which allows NIST computing DET plots, and an actual decision value (true/false), allowing C_{DET} and $\min C_{DET}$ values to be computed per participant. Male and female data is included, but no cross-trials are performed to avoid over-optimistic error rates.

Figure 5: *Faunagram*, or speaker-dependent performance plot, showing non-target (red) and target (blue) scores per speaker. Models are vertically sorted as a function of the speaker-dependent EER threshold.



Conversations were extracted from the English-spoken *Switchboard* corpus, a spontaneous conversational speech corpus consistent of thousands of telephone conversations from hundreds of speakers, each conversation typically running through 5 minutes. In 2000 and 2001, the Spanish-spoken spontaneous but not conversational *Ahumada* database (Ortega-Garcia, Gonzalez-Rodriguez, & Marrero-Aguilar, 2000) was explored with similar same-language (in train and test) protocols. Different releases of Switchboard were produced through the years (Cieri, Miller, & Walker, 2003), including extensive landline (local- and long-distance) and cellular (GSM and CDMA) phone data. Different telephone numbers and types of telephone handsets were used by the speakers and explored in the evaluations, showing the degradation on performance in mismatched conditions if not properly addressed. Systems were tested with one- or two-session training, significantly reducing the sensitivity to session mismatch by including explicit session variability in the model with the use of multiple session training. The effect of the length of the test was also explored, showing very limited benefit from same-session audio segments longer than 30-60 seconds, but strong degradation for shorter durations. However, with just 3 seconds of speech, systems were still providing very useful information (e.g., degradations

from 6% of EER with 30 seconds of test speech, to 15% of EER with just 3 seconds of speech). To give an idea of the complexity of each task, the cellular evaluation of 2001 involved 190 speakers and 2200 test speech segments, resulting in thousands of cross-speaker trials. For interesting details on the early evaluation campaigns and summary of systems and results in different conditions, readers can refer to (Doddington, Przybocki, Martin, & Reynolds, 2000).

Different additional tasks were deeply explored besides one-speaker detection. *Two-speaker detection* is essentially the same as one-speaker, but the test conversation includes both speakers (summed channels) in the conversation. A two-speaker mode for model training was also explored, where a target speaker was present in three different summed channel conversations with three different speakers. The task of *speaker segmentation* (later called *diarization*) has also been largely explored, which is the task of determining the time intervals during which unknown speakers are actually speaking in different datasets from telephone conversational speech, broadcast news and recordings of multi-speaker meetings. *Speaker tracking* was also explored from 1999 to 2001, being the task of determining the time instants where a known speaker was actually speaking in a multispeaker conversation. *Unsupervised adaptation*, where trials are presented sequentially and accepted ones can be used to further improve the model for future trials, was also explored from 2004 to 2006. Finally, in 2010 and 2012, a *Human Assisted Speaker Recognition* (HASR) task was proposed, where any combinations of machines, naive listeners or human experts were allowed in order to perform speaker detection over a manual selection of “especially difficult” trials from the core condition (Greenberg et al., 2011).

4.2. Challenges for GMM-UBM systems

In order to tackle all the sources of variability described above, short-term spectral systems had to be improved significantly. Cross-condition trials with handset, telephone number, and landline/mobile variability stressed the need for robust “channel-independent” features. Different channels show different long-term frequency responses, severely influencing with a cepstral additive component throughout the utterance to the short-term spectral estimates of the speaker. Cepstral mean (and variance) subtraction (Furui, 1981) proved useful in reducing channel effects. RASTA band-pass filters (Hermansky & Morgan, 1994), which additionally filter out frequency modulations not expected from speech, also helped to improve system performance. But channels are not strictly time-invariant, and time-dependent feature warping (Pelecanos & Sridharan, 2001) in sliding windows of 3 seconds significantly contributed to the robustness of the systems, making CMN-RASTA-Warping a by-the-time standard front end.

However, pooling together the scores from all target and non-target trials in order to compute a single DET plot, EER value or C_{DET} for a given system arose the problem of score misalignment. When we talked before about the target and non-target score distributions, we would like to observe similar distributions for all speakers (assuming gaussianity this means similar means and variances). However, it is well known that the *fauna* of speakers is varied (Doddington, Liggett, Martin, Przybocki, & Reynolds, 1998). While most speakers in a experiment behave similarly (*sheep*), some of the speakers are difficult to be correctly recognized (*goats*, target trials with low scores), some of the speakers are easy to be impostor (*lambs*, or speaker models giving high scores in non-target trials), and some of the speakers are successful impostor other speakers (*wolves*, speakers in non-target trials with high scores when accessing other speaker models). This behavior is illustrated in figure 5 (we called that plot a *faunagram*), where each horizontal line represents actual non-target and target scores for a given model, and models are vertically sorted as a function of their speaker-dependent EER threshold, clearly showing that almost all speakers show different means and variances of both target and non-target distributions. Even though the discrimination per speaker was reasonable (for instance, a low average of the EERs per speaker), for any single global threshold we select, the global EER will be significantly lower, as some speakers will be favored but some others will be strongly penalized.

In order to have a “common” non-target distribution for all speakers, scores are usually Z-normalized. This technique, known as *Z-norm*, estimates the distribution (mean and variance) of non-target scores of a given model using an external cohort of impostor trials, usually formed by other utterances from speakers different from the target, ideally in conditions similar to those of the testing environment. Then, all trial scores from this model are normalized to a zero mean unity variance Gaussian simply subtracting its speaker-dependent mean and dividing by its standard deviation. Once all speakers share a common impostor distribution, a single global threshold can be set giving a global EER close to the average of the EERs per speaker.

As cohort speakers for Z-norm are selected from the available data in the development phase, there is always some mismatch between their scores and the actual non-target distribution in the test phase, so some residual misalignment is always present. Moreover, different channels or handsets produce different non-target distributions (one per channel/handset). This is why *H-norm* (Handset normalization) was proposed, which is a double version of Z-norm, one per handset (carbon/button). During a test-time a decision about the estimated handset is needed in order to use the proper handset parameters for normalization. This is done simply scoring the input utterance against two UBMs, one per handset, and selecting the one giving the highest score (Reynolds et al., 2000). However, for multiple cross channel conditions,

the large number of combined “channel” models would make unlikely for them to be properly identified, so more sophisticated approaches to channel compensation are needed, as will be shown in section 6.2.

A second approach to score normalization known as *T-norm* became extremely popular (Auckenthaler, Carey, & Lloyd-Thomas, 2000). In this case, the non-target score distribution is estimated with the actual test speech against a cohort of (external) speaker models. This normalization is extremely efficient as it takes also into account the length of the input speech, acting also as a kind of test-duration normalization. Moreover, *T-norm* usually had the effect of a counter-clock-wise tilt of the DET curve. Then, as the usual application parameters in NIST evaluations for one-speaker detection resulted in upper-left (in the DET plot) desired operation points, *T-norm* not only benefited the discrimination but had an enormous benefit for the calibration of the system, allowing for detection thresholds resulting in much lower C_{DET} values. The joint use in cascade of *Z-* and *T-* norm, known as *ZT-norm* has become a standard score normalization technique that systems have successfully used for more than a decade.

5. GOING HIGHER (2002-2005): SUPRA-SEGMENTAL SYSTEMS

In the influential and pioneering work presented in (Doddington, 2001), word bi-grams were computed just from the word transcriptions provided by an ASR giving excellent speaker recognition results. Moreover, significant gains were reported from more and more training and testing data, as opposed to short-term spectral systems whose performance saturates beyond 30-60 seconds of test speech and two sessions for training. This success resulted in an explosion of approaches taking advantage of non-cepstral information. Using different speaker information extraction techniques, as shown in section 2.2, participants developed different nature systems with a common objective for exceptionally demanding new tasks and conditions as will be shown in Section 5.1.

5.1. Task definitions, corpora and evaluation conditions

While early evaluations focused on telephone channel (landline versus cellular), handset (carbon vs. electret) and duration effects, NIST evaluations since 2001 included new data types and conditions, and since 2004 through the use of the new Mixer corpus, which broadened the scope of evaluation of speaker recognition systems, the evaluation data allowed for large multi-session training, multi-microphone recording and multilingual speakers (bilingual speakers of English and a second language).

The *extended data* task allowed for multi-session training of speakers, providing up to 16 conversations

to train every speaker. Training conditions were explored with speech lengths of 10 seconds (from one conversation side), 1 full conversation side (average 2.5 minutes of speech), 3 sides (average 7.5 minutes of speech), 8 sides (20 minutes of speech) and 16 sides (40 minutes of speech). Those conditions provided enough data for non-segmental speaker recognizers requiring longer speech segments to fully exploit the different prosodic, ASR-conditioned and idiolectal high level systems. Different test segment lengths have also been explored, namely 10 seconds, 30 seconds, 1 side (average 150 seconds) and 1 summed-channel conversation (5 minutes from two speakers). Short-term cepstral systems had then the opportunity to focus on the demanding 10s-10s condition or the regular 1side-30 seconds, while high-level systems focused in the much bigger 8- or 16-sides training tasks.

Moreover, *multichannel microphone data* was obtained from hundreds of speakers who made some of their calls from one of three special on-site recording rooms where simultaneous recording of the phone conversations were obtained from eight different microphones:

- Ear-bud/lapel mike
- Mini-boom mike
- Courtroom mike
- Conference room mike
- Distant mike
- Near-field mike
- PC stand mike
- Micro-cassette mike

For instance, in tests performed with telephone-only data training, cross-microphone tests trials showed degradation from 2% EER for telephone test speech, to EER values from 4% to 8% depending on the microphone type, doubling or quadrupling the error rates. This multi-microphone data allowed for extensive testing of cross-channel conditions, converting the “discrete” previous channel variability (e.g., two handsets, or two types of telephone connections, etc.) in a “continuous” source of variation due to the large number of combinations of microphones, handsets, telephone channels and train/test durations, motivating a new “continuous” approach to session and channel variability, as will be shown in 5.3 and will explode as the core technique to state-of-the-art speaker recognition in section 6.2.

The *Mixer corpora*, in order to explore language effects on performance, have included hundreds of fluent bilingual speakers in English and a second language, namely (numbers are given for SRE 2004):

- Arabic (52 speakers)
- Mandarin (46 speakers)
- Russian (48 speakers)
- Spanish (79 speakers)
- English only (85 speakers)

Those 310 target speakers in SRE 2004 allowed checking that for matched language trials results were

mostly independent of the chosen language. However, for mismatched trials (the speaker uses different languages for train and test), a significant but not dramatic drop of performance was observed (e.g., from 3% EER for matched language trials to 8% EER for mismatched language trials).

In order to give an idea of the computational complexity of those evaluations, NIST departed from some thousands of trials in the early evaluations to about 26,000 trials (comparisons) involving over 400 speaker models in the SRE 2004 1side-1side task, or 80,000 trials from 1,900 speaker models (same speakers in different training conditions) for the SRE 2004 multisession (16-, 8-, 3- and 1-side) training trials.

5.2. Higher-level systems

In those years two different approaches were developed to further advance the current performance of speaker recognition systems in the presence of such challenging channel and language variability. The first approach, described in this section, took advantage of the much larger amounts of data and explicit session variability included for training in the extended data tasks with up to 16 full conversation sides, which were used for training prosodic and phone- and lexical-sequence based speaker models. The second approach, described in Section 5.3, faced channel and session variability through new cepstral-based ways of representing an utterance in a high-dimensional space, allowing for much better results in classical 1side-1side and shorter tasks as 1side-10sec or 10sec-10sec.

The different *high-level approaches* that were adopted can be exemplified from those reported in the 6-week summer workshop in 2002 at John Hopkins University (Reynolds et al., 2003). The 2001 NIST extended data was the reference task which used the entire Switchboard I conversational telephone speech database, with models trained up to 16 full conversation sides (about 40 minutes of speech), having close to 500 speakers and 4100 models (different models for different amounts of training data) and 57,000 test trials involved. To summarize, several families of systems were developed (results are provided for the 8-conversation training condition):

- *Acoustic*: a 2048 mixture cepstral-based GMM-UBM built from Switchboard II external data was used as reference system, providing an EER of only 0.7%.
- *Prosodic*: pitch and energy distributions and dynamics, through the joint slope modeling of pitch and energy contours gave an EER of 9.2%, which dropped to 5.2% when adding phone context to duration and contour dynamics. Additionally, a different system using 11 duration-derived statistics and 8 pitch related statistics obtained an EER of 8.1%.

- *Phone-sequence*: the idea of those systems is to exploit the information provided by simultaneous multiple open-loop phone recognizers in different languages. An open-loop recognizer is basically a speech recognition system without any language modeling, providing just the most likely sequence of phones without any linguistic constraint. Current open-loop phone recognizers have very high phone error rates (usually bigger than 30-40%), but they are expected to produce “speaker-dependent” transcriptions, as they are expected to err consistently for a given speaker. The combined use of phone n-grams from 5 different speech recognizers in 5 different languages (PPRLM) obtained an EER of 4.8% in the 8-conversation reference condition, while binary trees with a 3 token history (equivalent to 4-grams) obtained a 3.3% EER in the same task. Another system exploited the cross-stream information from the multiple phone streams, obtaining a 4.0% EER, which fused with the PPRLM was reduced to 3.6%.
- *Pronunciation modeling*: comparing constrained word-level ASR phone streams (the “true” phone sequence) with error-prone open-loop phone streams, speaker-dependent pronunciations were learned, obtaining an EER of 2.3%.
- *Lexical features*: n-gram idiolect systems as those described above in Doddington (2001) were tested, providing an EER of 11% from the best-available ASR word transcriptions.
- *Conversational features*: feature vectors were derived from turn-taking patterns and conversational style based information from pitch and durations, then converted into n-grams which obtained an EER of 15.2%.

Fusion of classifiers is strongly benefited from minimum correlation between systems to be fused. High-level systems produced excellent error rates with extremely different features and models from those in short-term cepstral-based GMM-UBM, which represented by the time the state-of-the-art of speaker recognition. When the non-cepstral systems were fused in the 8-conversation train task, the fused EER was exactly the same (0.7%) of the cepstral GMM-UBM system. Moreover, when all high-level and acoustic systems were combined, the reported EER was only 0.2%, a 71% relative reduction from incorporating high-level knowledge to a reference cepstral-based system. A similar combination of cepstral and high-level systems in an actual submission to NIST SRE 2004 is described in Kajarekar (2005).

5.3. High-dimensionality spectral systems

The GMM-UBM is a generative modeling approach, where the underlying model (the GMM) is supposed to be “generating” the observed features. In the late 90’s, a new discriminative pattern recognition technique

known as *Support Vector Machine* (SVM) (Schölkopf & Smola, 2002; Vapnik, 1995) showed to be extremely efficient discriminating objects (points) in very high dimensional feature spaces. SVMs are easily trained, the SVM speaker model being the hyperplane separating the target training speaker utterances (one or several points) from all non-target speaker utterances (lots of points). The score from an unknown utterance with respect to a given SVM model is obtained as a signed distance to the separating hyperplane. But in order to use SVM, low-dimensional feature vectors (as cepstral vectors or n-gram probabilities) must be transformed into separable high-dimensional vectors, which is done by *kernels* (or transformations).

Different kernel-tricks were proposed to transform a speech utterance, represented usually as a sequence of observed feature vectors, into a very high dimensional space representation. While GLDS (Campbell, Campbell, Reynolds, Singer, & Torres-Carrasquillo, 2006a) or MLLR (Stolcke, Kajarekar, Ferrer, & Shrinberg, 2007) supervectors were also successful, we highlight here the Gaussian (or GMM) supervector (GSV) concept (Campbell, Sturim, & Reynolds, 2006b), being a natural extension of the well-known GMM-UBM framework. A GMM of M Gaussian mixtures in P feature dimensional space (cepstral vectors of dimension P) is composed of M weights, $M \times P$ means and $M \times P$ variances (assuming diagonal covariance matrices). But as MAP adaptation of the GMM from the UBM is usually performed just over the means (weights and covariance matrices are shared across all speakers), one UBM-adapted speaker GMM differs from another GMM of a different speaker just in their means. A supervector is then just the stacked pile of all GMM means (after variance normalization), a vector of $M \times P$ values, or a single point in an $M \times P$ dimensional space. As typical values of M are 1024 or 2048, and P takes values from 20 to 60, speech utterances will then be represented by vectors of size 20k to 120k, in other words, by points in a 20k-120k dimensional space.

Once utterances are reduced to points in a high-dimensional space, the problem of *session variability* (between-session differences in the speaker information) becomes variability between points, with some dimensions (directions) more severely affected than others. The problem of session variability compensation can be then addressed estimating the principal directions of channel variability in the development phase, and later canceling them out in the test phase, a process known as *Nuisance Attribute Projection* (NAP) (Solomonoff, Campbell, & Boardman, 2005). If we compute for every speaker in the development set the mean supervector (one per speaker), and every utterance is normalized subtracting its mean speaker vector, the resulting data set (known as *within-scatter* matrix) with all normalized utterances contains only session variability and no speaker information. Then, an eigenvector analysis (PCA, Principal Components Analysis) of this within-scatter matrix will provide the desired principal direc-

tions of channel variability, known as *eigenchannels*. In the test phase, we can easily project every unknown supervector into those channel dimensions, and subtract the resulting supervector (which should contain the session variability components) from the original one. In order to get an idea of the significant improvements obtained with GSV-SVM and NAP facing channel variability, readers are referred to Campbell et al. (2006b).

6. BIG DATA EVALUATIONS (2006-2012): SESSION VARIABILITY COMPENSATION

Since 2006, NIST SREs became biannual, and have introduced significant changes evaluation after evaluation, as shown below in 6.1, additionally introducing massive amounts of new data in every new evaluation. But the biggest difference, as highlighted in Brümmer et al. (2007), is that especially from SRE 2006, “systems no longer train individual speaker models from some minutes of speech, but whole systems are trained on hundreds of hours of speech in whole NIST SRE databases” (p. 2082), transforming the conceptually simple speaker detection task, classically seen as that of comparing two utterances to determine if they come or not from the same speaker, into a serious *big data* task where systems are designed to jointly optimize the detection of thousands of speakers in hundreds of thousands of comparisons, where the speech segments in the comparisons are tens of thousands of utterances of varied and mixed channel, speaking style, duration and noise characteristics.

6.1. Task definitions, corpora and evaluation conditions

The last four evaluations in the NIST SRE series have introduced major changes from evaluation to evaluation that can be summarized as follows:

- **SRE 2006:** the eight *alternate microphones* from Mixer 3, as shown in section 5.1, were fully exploited. Additionally, an *alternative cost function*, C_{llr} , is included as optional but soon became widely used being usually the objective function for cost minimization in fusion of systems. The number of trials in the required (*Iside-Iside*) condition was about 54,000.
- **SRE 2008:** up to 2006, evaluations dealt with spontaneous conversational speech obtained in telephone conversations between remote speakers. The *Phoncall* conversational speech database, known as Mixer 3, was used for SRE 2008. Additionally to regular telephone recordings (*phoncall-phn*), conversational telephone speech recorded over a microphone channel (*phoncall-mic*) is also included in the test conditions. But for SRE 2008, a new type of speech was recorded in an on-site

interview scenario, resulting in the Mixer 5 Interview speech database. It is also conversational speech but of a totally different nature, and it is recorded via multiple simultaneous microphones (*interview-mic*). In the required condition (known as *short2-short3*, similar to a multichannel *Iside-Iside*), 1,788 phonecall and 1,475 interview speaker models were involved for a total of almost 100,000 trials (60,000 male and 40,000 female).

- **SRE 2010:** four new significant changes were introduced in 2010. Firstly, all speech in the evaluation was *English* (both native and non-native). Second, some of the conversational telephone speech data has been collected in a manner to produce particularly high or particularly low *vocal effort*. Third, the interview segments in the required condition were of *varying duration*, ranging from three to fifteen minutes. And finally, the most radical change, a *new performance measure* ($\text{new_C}_{\text{DET}}$), intended for systems to obtain good calibration at extremely low false alarm rates, was defined from a new set of parameter values. While C_{DET} always used $\text{C}_{\text{miss}}=10$, $\text{C}_{\text{FA}}=1$ and $\text{P}_{\text{target}}=0.01$, the new C_{DET} parameters are $\text{C}_{\text{miss}}=1$, $\text{C}_{\text{FA}}=1$ and $\text{P}_{\text{target}}=0.001$, increasing in a factor of 100 the relative weight of false alarms to missed detections. In order to have statistically significant error rates for very low false alarm rates, the required (core) condition included close to 6,000 speaker models, 25,000 test speech segments and 750,000 trials.
- **SRE 2012:** in NIST's own wording, "SRE12 task conditions represent a significant departure from previous NIST SRE's (NIST 2012, p. 1)". In all previous evaluations, participants were told what speech segments should be used to build a given speaker model. However, in 2012, all the speech from previous evaluations with known identities was made available to build the speaker models, resulting in availability of large (and variable) number of segments from the *phone-tel*, *phone-mic* and *int-mic* conditions, systems being free to build their models using whatever combination of previous files from every speaker. Moreover, additive and environmental noises were included, variable test durations (300, 100 and 30 seconds) considered, knowledge of all targets was allowed in computing each trial detection score, tests were performed with both known and unknown impostors, and for the first time in a long time ASR word transcripts were not provided. Finally, the cost measure was changed again, averaging the 2010 operating point (optimized for very low false alarm rates) with a new one with a greater target prior (pushing the optimal threshold back, closer to its "classical" position), intending for systems to show greater stability of the cost measure and good score calibration over a wider range of log-likelihoods. For this evaluation, close to 2,250 target speakers and 100,000 test segments were involved, for a total

in the required (core) condition of 1,381,603 trials. For those involved in the (optional) extended trials task, the number of trials was 67,000,000.

Factor analysis and i-vectors

Even though one site continued to submit successful high- and low-level combined systems in those big data evaluations (Ferrer et al., 2013; Kajarekar et al., 2009; Scheffer et al., 2011), there was a consensus in turning back to cepstral-only systems. The computational complexity of higher-level systems and the relative improvements obtained in limited training data conditions helped the community to move towards a scientifically complex but very rewarding approach because of the performance and computational efficiency of new high-dimensional spectral systems as JFA-compensated GMM-UBM, and later, i-vector front-end extraction and PLDA based classification.

Different supervectors from different recordings of the same speaker show severe variability due to *intersession variability*, accounting for channel and speaker specific variability. In order for a test supervector to be close to the target speaker one, intersession variability, usually called just "channel" variability, must be compensated. Joint Factor Analysis (JFA) (Kenny, 2005) models channel variability explicitly, taking the variability of a supervector as a linear combination of the speaker and channel components. In order to know and compensate the channel "offset" in a test utterance, the main directions of channel and speaker variability in the high-dimensional space have to be found in advance from large development datasets. The eigenchannels matrix can be initialized through PCA of the within-scatter matrix as shown in 5.3, and the *eigenspeakers* (also called *eigenvoices*) one in a similar way from the *between-scatter* matrix, that is, a data structure with one column per speaker, where every column is the speaker mean vector (mean of the different session-dependent speaker supervectors) minus the global mean of all speaker means. After this PCA initialization of both matrices, they are improved through several Expectation-Maximization (EM) iterations over the whole development dataset. Once the eigenchannel and eigenspeaker matrices are estimated, the channel and speaker factors in a test utterance are jointly estimated as point estimates as in classic relevance MAP. Then the channel factor can be discarded, and the "clean" speaker supervector, estimated as the offset from the UBM supervector in an "amount" given by the speaker factor in the eigenspeaker directions, can be used for recognition with a synthesized "clean" GMM model or a SVM with "clean" supervectors. JFA-based approaches, in several of the numerous flavours of this technology, have obtained excellent results in the 2006 to 2010 NIST SREs.

However, it was shown that the channel factors, which are to be discarded from the model, still contain information from the speaker. Then, instead of assuming

two different variability subspaces (speaker and channel), in (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011) a single subspace is considered, called Total (T) Variability subspace, which contains both speaker information and channel variability. In this case, the T low-rank matrix is obtained in a similar way to the eigen-speakers matrix but all the utterances for each speaker are included (as if each of them came from different speakers). Once the T matrix is available, an UBM is used to collect Baum-Welch first-order statistics from the utterance. The high-dimensional supervector (dimensions ranging from 20,000 to 150,000), which is built stacking together those first order statistics for each mixture component, is then projected into a low dimensional fixed-length representation known as the i-vector in the subspace defined by T, being estimated as a MAP point-estimate of a posterior distribution.

As both target (train) and test utterances are now represented by fixed-length i-vectors (typically from 200 to 600 dimensions), it can be seen as a new “global” feature extractor, which captures all relevant speaker and channel variability in a given utterance into a low-dimensional vector. Target and test i-vectors can be directly compared through cosine scoring, a measure of the angle between i-vectors, which produce recognition results similar to those with JFA. But as they still contain both speaker and channel information, factor analysis can also be applied in the total variability subspace to better separate low dimensional contributions of channel and speaker information. Probabilistic Linear Discriminant Analysis (PLDA) (Prince & Elder, 2007) models the underlying distribution of the speaker and channel components of the i-vectors in a generative framework where Factor Analysis is applied to describe the i-vector generation process. In this PLDA framework, a likelihood ratio score of the same speaker hypothesis versus the different speaker hypothesis can be efficiently computed, as a closed form solution exists. Further, in order to reduce the non-Gaussian behavior of speaker and channel effects in the i-vector representation, i-vector length normalization (Garcia-Romero, & Espy-Wilson, 2011) is proposed allowing the use of probabilistic models with Gaussian assumptions as PLDA instead of complex Heavy Tailed representations.

I-vector extraction, PLDA modeling and scoring, and i-vector length normalization have become by the time of writing this paper the current state-of-the-art in text-independent speaker recognition, and the basis of successful submissions to NIST SRE 2012. Multiple generative and discriminative variants, optimizations and combinations of the above ideas exist based in the same underlying principles. As a result, joint submissions to NIST SREs of multiple systems from multiple sites into a single fused system have become usual, and a must if an individual system wants to be in the horse-race photo-finish of best submissions (Saeidi et al., 2013).

However, recent success of Deep Neural Networks in different areas of speech processing (Hinton et al.,

2012; Lopez-Moreno et al., 2014) promise for the near future exciting developments in speaker recognition, as those advanced in Vasilakakis, Cumani, and Laface (2013), and Variani, Lei, McDermott, Lopez-Moreno, and Gonzalez-Dominguez (2014).

7. DEMYSTIFYING SREs: THE 2014 NIST I-VECTOR CHALLENGE

As shown in the above sections, NIST Speaker Recognition Evaluations have always demanded from participants a very complex machinery of signal processing, pattern recognition, data engineering and computational resources. A newcomer to the evaluations receives hundreds of hours of speech data from previous evaluations as development data, with varied data structures and different segments and speaker identity labeling formats in a mixture of conditions (channels, speaking style, durations...). Even if having available and properly working all the software components to build a system, the human and computational resources to be spent for voice activity detection, feature extraction, universal background modeling, estimation of variability subspaces, etc., with the proper separation of conditions and tasks, is a major access obstacle that inhibits many potential participants from enrolling in the evaluations.

In order to eliminate this barrier and promote participation from pattern recognition scientists working in different areas, a simplified exercise has been proposed in 2014 consisting in the classification and recognition of a large amount of properly-extracted but unlabeled i-vectors (speaker identifiers are not provided). Moreover, a good i-vector cosine-scoring reference system is provided, with all the necessary scripts to work with the evaluation data and submit results of the evaluation. Additionally, NIST has made available an on-line cost scoring system (over 40% of the test data) that provides participants in real-time a good estimate of the goodness of every new algorithm or tuning factor they have tested. And finally, all participant-best (minimum) costs and associated submitting site names are known every time a participant submits a new system, promoting a horse-race competition where all participants see each other progress.

By the time of writing this paper, the participation level in the 2014 i-vector challenge is a major success, and the reported cost improvements over the reference system promise exciting news in the form of new or optimized algorithms to be presented in the NIST challenge workshop to be held during the Odyssey Speaker and Language Recognition conference in June 2014.

8. DISCUSSION AND CONCLUSION

The NIST series of Speaker Recognition Evaluations is a good example of how to foster tremendous progress in a specially challenging problem from a simultaneously

competitive and cooperative scientific community. Nevertheless, linking the progress in speaker recognition to participation in the NIST cycle of evaluations is not free of inconveniences.

First of all, participants cross-benefit from the other participants publications and previous submissions, resulting in a pseudo-normalization of procedures and system components, with plenty of submissions differing in very slight details. Moreover, high-risk innovation is indirectly penalized as new alternative approaches are unable to reach or help existing state-of-the-art systems, the task being so complex and mature that participants tend to bet on winning horses. And finally, being a competitive evaluation according to a given cost function with a final rank of participants in every task, and given the benefits of fusing multiple classifiers in a complex task like this, participants tend to group themselves in large consortia, both to minimize risks and opt to better ranking, submitting valid but unrealistic fusions of multiple (even tens of) systems.

Speaker recognition systems have shown extremely good performance and computational efficiency when lots of development data, in conditions close to the evaluation (application) data, are available. However, the current biggest challenge to speaker recognition is how to adapt this well-proven technology in known domains to new applications when little (compared to the hundreds of hours of data in NIST SREs) or no development data is available in new unknown conditions (language, channels, speaking style, audio quality, etc.). Related challenging issues are those of calibration in highly mismatched environments, the production of reliable automatic recognition results from descriptive speech features (pronunciation patterns, voicequality, prosodic features...) correlating with linguists and phoneticians observations, or measuring the goodness of the system decisions in individual comparisons, that is, how reliable is one system in the unknown comparison at hand, not globally in thousands of known comparisons.

However, in spite of the inconveniences and out-of-domain limitations, it is extremely beneficial for system developers to be involved in the evaluations. There is a huge gap between developing a new system in the laboratory or trying a new pattern recognition algorithm and being able to obtain good results in the demanding conditions of the NIST evaluations. And only when tested in really challenging environments, speaker recognition systems will be a step closer to being usable in daily applications.

REFERENCES

- Adami, A. G., Mihaescu, R., Reynolds, D. A., & Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition. *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, 4, 788–791. <http://dx.doi.org/10.1109/ICASSP.2003.1202761>
- Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3), 42–54. <http://dx.doi.org/10.1006/dspr.1999.0360>
- Brümmer, N. (2010). *Measuring, refining and calibrating speaker and language information extracted from speech* (doctoral dissertation). University of Stellenbosch, South Africa.
- Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., ... Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7), 2072–2084. <http://dx.doi.org/10.1109/TASL.2007.902870>
- Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006a). Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2–3), 210–229. <http://dx.doi.org/10.1016/j.csl.2005.06.003>
- Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006b). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5), 308–311. <http://dx.doi.org/10.1109/LSP.2006.870086>
- Cieri, C., Miller, D., & Walker, K. (2003). From switchboard to fisher: Telephone collection protocols, their uses and yields. *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 – INTERSPEECH 2003*, 1597–1600.
- Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 357–366. <http://dx.doi.org/10.1109/TASSP.1980.1163420>
- Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014, May). *COVAREP - A collaborative voice analysis repository for speech technologies*. To be presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '14), Florence, Italy.
- Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798. <http://dx.doi.org/10.1109/TASL.2010.2064307>
- Doddington, G. R. (2001). Speaker recognition based on idiolectal differences between speakers. *Proceedings of the 7th European Conference on Speech Communication and Technology, EUROSPEECH 2001 – INTERSPEECH 2001*, 2521–2524.
- Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings of the International Conference on Spoken Language*, 1–5.
- Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A. (2000). The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective. *Speech Communication*, 31(2–3), 225–254. [http://dx.doi.org/10.1016/S0167-6393\(99\)00080-1](http://dx.doi.org/10.1016/S0167-6393(99)00080-1)
- Ferrer, L., McLaren, M., Scheffer, N., Lei, Y., Graciarena, M., & Mitra, V. (2013). *A noise-robust system for NIST 2012 speaker recognition evaluation*. Paper presented at the 14th INTERSPEECH Conference 2013, Lyon, France.
- Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 254–272. <http://dx.doi.org/10.1109/TASSP.1981.1163530>
- Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. *Proceedings of the 12th INTERSPEECH Conference 2011*, 249–252.
- Gonzalez-Rodriguez, J. (2011). Speaker recognition using temporal contours in linguistic units: The case of formant and formant-bandwidth trajectories. *Proceedings of the 12th INTERSPEECH Conference 2011*, 133–136.
- Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., & Ortega-Garcia, J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7), 2104–2115. <http://dx.doi.org/10.1109/TASL.2007.902747>

- Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G., & Godfrey, J. (2011). *Human assisted speaker recognition in NIST SRE10*. Paper presented at the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11), Prague, Czech Republic.
- Hébert, M. (2008). Text-dependent speaker recognition. In J. Benesty, M. Sondhi, & Y. Huang (Eds.), *Springer handbook of speech processing* (pp. 743–762). Berlin–Heidelberg, Germany: Springer. http://dx.doi.org/10.1007/978-3-540-49127-9_37
- Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4), 578–589. <http://dx.doi.org/10.1109/89.326616>
- Hernando, J., & Nadeu, C. (1998). Speaker verification on the polycost database using frequency filtered spectral energies. *Proceedings of the 5th International Conference on Spoken Language*, 98, 129–132.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6), 82–97. <http://dx.doi.org/10.1109/MSP.2012.2205597>
- Kajarekar, S. S., Ferrer, L., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A., & Zheng, J. (2005). SRI's 2004 NIST speaker recognition evaluation system. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 1, 173–176.
- Kajarekar, S. S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L., & Bocklet, T. (2009). THE SRI NIST 2008 speaker recognition evaluation system. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 4205–4208. <http://dx.doi.org/10.1109/ICASSP.2009.4960556>
- Kenny, P. (2005). *Joint factor analysis of speaker and session variability: Theory and algorithms* (Technical Report No. CRIM-06/08-13). Montreal, Canada: CRIM.
- Khoury, E., Vesnicer, B., Franco-Pedroso, J., Violato, R., Boulknafet, Z., Mazaira Fernandez, L. M., ... Marcel, S. (2013, June). The 2013 speaker recognition evaluation in mobile environment. *2013 International Conference on Biometrics (ICB)*, 1–8. <http://dx.doi.org/10.1109/ICB.2013.6613025>
- Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1), 12–40. <http://dx.doi.org/10.1016/j.specom.2009.08.009>
- Kockmann, M., Ferrer, L., Burget, L., Shriberg, E., & Černocký, J. (2011). Recent progress in prosodic speaker verification. *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, 4556–4559.
- Larcher, A., Aronowitz, H., Lee, K. A., & Kenny, P. (Organizers) (2014, September). *Text-dependent speaker verification with short utterances*. Special session to be conducted at the 15th INTERSPEECH Conference 2014, INTERSPEECH 2014 (Singapore). Retrieved from http://www.interspeech2014.org/public.php?page=special_sessions.html
- Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Communication*, 60, 56–77. <http://dx.doi.org/10.1016/j.specom.2014.03.001>
- Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez-González, D., Gonzalez-Rodriguez, J., Moreno, P.J. (2014, May). *Automatic language identification using deep neural networks*. To be presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '14), Florence, Italy.
- Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. *Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH 1997*, 1895–1898.
- National Institute of Standards and Technology (NIST) (2012). *The NIST year 2012 speaker recognition evaluation plan*, 1–7. Retrieved from http://www.nist.gov/itl/iad/mig/upload/NIST_SRE_evalplan-v11-r0.pdf
- Ortega-Garcia, J., Gonzalez-Rodriguez, J., & Marrero-Aguilar, V. (2000). AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. *Speech Communication*, 31(2), 255–264. [http://dx.doi.org/10.1016/S0167-6393\(99\)00081-3](http://dx.doi.org/10.1016/S0167-6393(99)00081-3)
- Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification. *Proceedings of 2001: A speaker odyssey: The speaker recognition workshop*, 213–218.
- Plumpe, M. D., Quatieri, T. F., & Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5), 569–586. <http://dx.doi.org/10.1109/89.7841090>
- Prince, S. J. D., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. *IEEE 11th International Conference on Computer Vision*, 1–8. <http://dx.doi.org/10.1109/ICCV.2007.4409052>
- Ramos, D., Gonzalez-Rodriguez, J., Zadora, G., & Aitken, C. (2013). Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58(6), 1503–1518. <http://dx.doi.org/10.1111/1556-4029.12233>
- Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., ... Xiang, B. (2003). The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, 4, 784–787.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41. <http://dx.doi.org/10.1006/dspr.1999.0361>
- Rose, P. (2002). *Forensic speaker identification*. CRC Press. <http://dx.doi.org/10.1201/9780203166369>
- Saeidi, R., Lee, K. A., Kinnunen, T., Hasan, T., Fauve, B., Bousquet, P. M., ... Ambikairajah, E. (2013, August). *14U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification*. Paper presented at the 14th INTERSPEECH Conference 2013, Lyon, France.
- Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E., & Stolcke, A. (2011). The SRI NIST 2010 speaker recognition evaluation system. *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, 5292–5295. <http://dx.doi.org/10.1109/ICASSP.2011.5947552>
- Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge MA: MIT Press.
- Shriberg, E. (2007). Higher-level features in speaker recognition. In C. Müller (Ed.), *Speaker classification I. Fundamentals, features and methods* (pp. 241–259). Berlin–Heidelberg, Germany: Springer. http://dx.doi.org/10.1007/978-3-540-74200-5_14
- Solomonoff, A., Campbell, W. M., & Boardman, I. (2005). Advances in channel compensation for SVM speaker recognition. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 1, 629–632. <http://dx.doi.org/10.1109/ICASSP.2005.1415192>
- Stolcke, A., Kajarekar, S. S., Ferrer, L., & Shrinberg, E. (2007). Speaker recognition with session variability normalization based on MLLR adaptation transforms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 15(7), 1987–1998. <http://dx.doi.org/10.1109/TASL.2007.902859>
- Thiruvaran, T., Ambikairajah, E., & Epps, J. (2008). FM features for automatic forensic speaker recognition. *Proceedings of the 9th INTERSPEECH Conference 2008*, 1497–1500.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York NY: Springer. <http://dx.doi.org/10.1007/978-1-4757-2440-0>
- Variani, E., Lei, X., McDermott, E., Lopez-Moreno, I., Gonzalez-Dominguez, J. (2014, May). *Deep neural networks for small footprint text-dependent speaker verification*. To be presented at the 2014 IEEE International Conference on

- Acoustics, Speech and Signal Processing (ICASSP '14), Florence, Italy.
- Vasilakakis, V., Cumani, S., & Laface, P. (2013, October). *Speaker recognition by means of Deep Belief Networks*. Technologies in Forensic Science, Nijmegen, The Netherlands.
- van Leeuwen, D. A., & Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In C. Müller (Ed.), *Speaker classification I* (pp. 330–353). Berlin–Heidelberg, Germany: Springer.