A Review of the Benefits and Issues of Speaker Verification Evaluation Campaigns

Asmaa El Hannani[1], **Jean Hennebert**[2,3]

[1]Deptartment of Computer Science, University of Sheffield, UK
[2]HES-SO, Business Information Systems, TechnoArk 3, CH-3960 Sierre, Switzerland
[3]University of Fribourg, Bd de Prolles 90, CH-1700 Fribourg, Switzerland
asmaa.elhannani@sheffield.ac.uk, jean.hennebert@hevs.ch

Abstract

Evaluating speaker verification algorithms on relevant speech corpora is a key issue for measuring the progress and discovering the remaining difficulties of speaker verification systems. A common evaluation framework is also a key point when comparing systems produced by different labs. The speech group of the National Institute of Standards and Technology (NIST) has been organizing evaluations of text-independent telephony speaker verification technologies since 1996, with an increasing success and number of participants over the years. These NIST evaluations have been recognized by the speaker verification scientific community as a key factor for the improvement of the algorithms over the last decade. However, these evaluations measure exclusively the effectiveness in term of performance of the systems, assuming some conditions of use that are sometimes far away from any real-life commercial context for telephony applications. Other important aspects of speaker verification systems are also ignored by such evaluations, such as the efficiency, the usability and the robustness of the systems against impostor attacks. In this paper we present a review of the current NIST speaker verification evaluation methods, trying to put objectively into evidence their current benefits and limitations. We also propose some concrete solutions for going beyond these limitations.

1. Introduction

Speaker verification consists in verifying a person's claimed identity. It is a subfield of speaker recognition that comprises all of the many different tasks of distinguishing people on the basis of their voices.

Speaker verification is also a subfield of biometric technologies. Biometrics, which bases the person authentication on the intrinsic aspects of a human being, appears as a viable alternative to more traditional approaches such as keys, badges, magnetic cards or memorized passwords. Biometric person authentication could be done using various modalities such as fingerprints, face, speech, dynamic signature, iris, hand geometry or keystroke dynamics. As a biometric modality, speech has a number of advantages and potentialities in comparison to the other modalities. Speech does not require any physical contact with the acquisition device and so is considered lowly intrusive by users. Moreover, in some cases (over the telephone, the radio, in the dark ...), speech is often the only available modality to recognize the identity of a person.

There are two main tasks of speaker recognition; speaker identification and speaker verification. The difference between these two tasks rests mainly on the type of decision that should be made. Usually, they are both based on the same modelling technologies (Wan and Campbell, 2000; Reynolds, 1995). The **speaker identification** task consists in determining, from a sequence of speech samples, the identity of an unknown person among N recorded speakers, called reference speakers. The identification answers the question "Whose voice is this?". This process gives place to N possible results. The **speaker verification** (also referred as speaker detection) aims to determine if a person, who claims to be a target speaker[1], is or is not this speaker. The decision will be either an acceptance or a rejection. The verification answers the question, "Am I who I claim to be?". If the person is not a target speaker, he is called an impostor.

Evaluating speaker recognition algorithms on relevant speech corpora is a key issue for measuring the progress and assessing the difficulties of speaker verification systems. In an ongoing effort to support research in text-independent speaker recognition technologies, NIST has been conducting annual speaker recognition evaluations. The aim of these evaluations is to provide common framework (data, rules and scoring) to allow focused technology development and meaningful comparison of techniques and approaches. However, these evaluations measure exclusively the effectiveness in term of performance of the systems, assuming some conditions of use that are sometimes far away from any real-life commercial context for telephony applications. Other important aspects of speaker verification systems are also ignored by such evaluations, such as the efficiency, the usability and the robustness of the systems against impostor attacks. Finally, using a similar evaluation framework for consecutive years have made converge many labs into using similar kind of algorithms (mostly GMM based) where system differences are essentially linked to the ability of the labs to aggregate large quantity of data used for training normalization and background components of the system.

2. Speaker Verification Evaluation

2.1. Performance Factors

Speaker verification performance is dependent upon many different factors that could be grouped in the following categories:

[1]Also referred in the literature as true, reference or client speaker

Proceedings of the ELRA Workshop on Evaluation at LREC 08, Marrakech, Morocco. 2008

- **Intra-speaker Variabilities**: Usually the speaker model is obtained using a limited amount of speech data that characterizes the speaker at a given time and situation. However, the voice can change in time due to aging, illness, emotions, tiredness and potentially other factors. For these reasons, the speaker model may not be representative of the speaker in all his/her potential states. Variabilities may not all be covered, which affect negatively the performance of the speaker verification systems. To deal with this problem, incremental enrollment techniques can be used in order to include the short and long-term evolution of the voice (see for example (Barras et al., 2004)).

- **Mismatch Factors**: The mismatch in recording conditions between the training and testing is the main challenge for automatic speaker recognition, specially when the speech signal is acquired on telephone lines. Differences in the background noise, in the telephone handset, in the transmission channel and in the recording devices can, indeed, introduce variabilities over the recording and decrease the accuracy of the system. This is mainly due to the statistical models that do not capture only the speaker characteristics but also the environmental ones. Hence, the system decision may be biased if the verification environment is different from the enrollment. The features and score normalization techniques (e.g. (Pelecanos and Sridharan, 2001; Reynolds et al., 2003; Auckenthaler et al., 2000; Reynolds et al., 2000)) are useful to make speaker modelling more robust to recording conditions. The high-level features (e.g. (Reynolds et al., 2003; Campbell et al., 2003; El Hannani and Petrovska-Delacrétaz, 2007)) are also important because they are supposed to be more robust to mismatched conditions.

- **Amount of Speech Data**: The amount of training data available to build the speakers model and to test it has also a large impact on the accuracy of the systems. This was confirmed during the NIST Speaker Recognition Evaluation (SRE) evaluations (Martin and Przybocki, 2004), where it has been shown that the duration and number of sessions of enrollment and verification affect the performance of the speaker verification systems.

2.2. Performance Measures

The performance of any speaker recognition system is evaluated in function of the error rate. There are two types of errors that occur in a verification task: the false acceptance when the system accepts an impostor and the false rejection when the system rejects a valid speaker. Both types of errors depend on the decision threshold. With a high threshold, the system will be highly secured. In other words, the system will make very few false acceptances but a lot of false rejections. If the threshold is fixed to a low value, the system will be more convenient to the users making few false rejections and lots of false acceptances. The rates of false acceptance, R_{FA}, and false rejection, R_{FR}, are then functions of the threshold and define the operating point of the system. They are calculated as follows:

$$R_{FA} = \frac{number\ of\ false\ acceptances}{number\ of\ impostors\ access} \quad (1)$$

$$R_{FR} = \frac{number\ of\ false\ rejections}{number\ of\ targets\ access} \quad (2)$$

These rates are normally estimated on the development set and are further used to compute the Detection Cost Function (DCF). This cost function is a weighted measure of both false acceptance and false rejection rates:

$$DCF = C_{FR}P_{tar}R_{FR} + C_{FA}P_{imp}R_{FA} \quad (3)$$

where C_{FR} is the cost of false rejection, C_{FA} is the cost of false acceptance, P_{tar} is the a priori probability of targets and P_{imp} is the a priori probability of impostors.
The DCF is the most used measure to evaluate the performances of operational speaker verification systems. The smaller is the value of the DCF, the better is the system for the given application and conditions. Thus, the decision threshold is usually optimized in order to minimize the DCF. This optimization is often done during the development of the system on a limited set of data.
Another popular measure is the Equal Error Rates (EER). It represents the error at the threshold which gives equal false acceptances and false rejections rates. The EER is not interpretable in function of the cost but still widely used as a reference indication of the performance of the system.

2.3. Detection Error Tradeoff Curve

The measures presented before evaluate the performances of the system in a single operating point. However, representing the performance of the speaker verification system over the whole range of operating points is also useful and can be achieved by using a performance curve. The Detection Error Tradeoff (DET) curve (Martin et al., 1997), a variant of the Receiver Operating Characteristic (ROC) curve (Egan, 1975), has been widely used for this purpose. In the DET curve the R_{FA} is plotted as a function of the R_{FR} and the axis follow a normal deviate scale. The points of the DET curve are obtained by varying the threshold T. This representation allows an easy comparison of the performances of the systems at different operating points. The EER appears directly on this curve as the intersection of the DET curve with the first bisectrix.

2.4. Speech Corpora and benchmarks

There has been a plethora of speaker verification algorithms and technologies proposed by the scientific communities and commercial vendors. Evaluating speaker recognition algorithms on relevant speech corpora has become a key factor for measuring the progress and detecting difficulties of speaker recognition systems. A survey of standard speech corpora that are suitable for the development and evaluation of speaker recognition systems can be found in (Godfrey et al., 1994; Campbell and Reynolds, 1999). The main suppliers of these corpora are the European Language Resources Association (ELRA)[2], the Linguistic Data

[2]http://www.elra.info

Consortium (LDC)[3], and the Oregon Graduate Institute (OGI)[4]. The most used corpora for speaker recognition are listed in Table 1.

Corpora	Supplier
SIVA PolyVar POLYCOST	ELRA
Switchboard I & II & Cellular TIMIT & NTIMIT & HTIMIT & CTIMIT NIST SREs Subsets Fisher KING YOHO SPIDRE CSLU TSID	LDC
Speaker Recognition Corpus	OGI

Table 1: Speaker recognition corpora and their suppliers.

Methodologies to benchmark the many different speaker recognition approaches have soon been developed on top of the available corpora. Generally speaking, one can classify benchmark methodologies as explained in (Cappelli et al., 2006) and as illustrated on Figure 2.4.:

- **In-house evaluation with self-defined test**: The testing protocol is self-defined on a privately owned database. Often, the recording conditions are controlled by the lab. As a consequence, results are not easily reproducible by a third party. The door is also open to data manipulation such as selection of speakers, discarding of outliers, etc. From an algorithmic point of view, problems of over-fitting the speaker data may also arise, i.e. the algorithms become too specific to a given data set. This is especially true if the evaluation protocol is not organized into independent development and evaluation sets.

- **In-house evaluation with existing benchmark**: The testing protocol and the corpora are publicly available. Assuming that the pre-defined evaluation protocols are strictly followed, results of algorithms executed on the data are comparable across sites and publications. Some existing corpora provide a defined evaluation procedure such as TIMIT (Reynolds et al., 1995), POLYCOST (Melin and Lindberg, 1996)(Hennebert et al., 2000), KING (Reynolds, 1994) or YOHO (Campbell, 1995). However, the risk of overfitting is definitively remaining as the protocols are often not organized into independent development and evaluation sets.

- **Independent weakly supervised evaluation**: The testing protocol is defined by an independent institute and the data, supposedly unseen by the participants, are made available just before the beginning of the test. Data samples are unlabeled (no ground truth) and the participant provide the evaluator with the results of the algorithms within given time constraints. All NIST Speaker Recognition Evaluations[5] fall in this category. For NIST, the unseen data are provided to the participants in a pretty large quantity to minimize the risk of participants willing to listen to the waveforms and manually perform the verification. However, the quantity of data is so large the time constraints are so strict that participating to such evaluations requires resources (human and cpu) that are often not available in participating labs.

- **Independent supervised evaluation**: The data are here completely sequestered by the evaluator. The participant provides the evaluator with a full solution to run the tests, including hardware and software. The evaluator can then better control the evaluation and the risk of human intervention is minimized. The drawback of the approach is that the evaluation can only be performed in terms of accuracy and not in terms of cpu or memory footprint.

- **Independent strongly supervised evaluation**: The participant provides here the evaluator with a software only solution that is run on the evaluator hardware. Recently, the 2007 Biometric Multimodal Evaluation Campaign was organized following the independent strongly supervised evaluation scheme described above. Speaker verification was present as part of the talking face evaluation (Fauve et al., 2008). Keeping the same hardware allows performing full comparisons in terms of performance, cpu and memory footprint. However, the drawback lies in the extra difficulty for the participant to modify its software so that it complies with a given input-output framework. The costs in terms of time and resources are also much larger on the side of the evaluator.

3. Overview of NIST Speaker Recognition Evaluation

3.1. History

The speech group of the National Institute of Standards and Technology (NIST) has now been organizing evaluations of speaker recognition technologies since 1996 with an increasing success over the years. The NIST Speaker Recognition Evaluation (SRE) campaigns varied from 1996 to 2006 in term of tasks and corpora used (Martin and Przybocki, 2004; Przybocki et al., 2006). The speaker detection (verification) task has remained the primary task over the years. However the evaluations have started including some other tasks such as speaker tracking and speaker segmentation. NIST included speaker tracking task between 1999 and 2001 and the speaker segmentation task between 2000 and 2002. The dadasets used for the evaluation have also changed to include different handsets, transmission types and languages. Table 2 summarizes the evolution of NIST SRE regarding the corpus, tasks, and training/testing durations.

[3] http://www.ldc.upenn.edu

[4] http://cslu.cse.ogi.edu

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

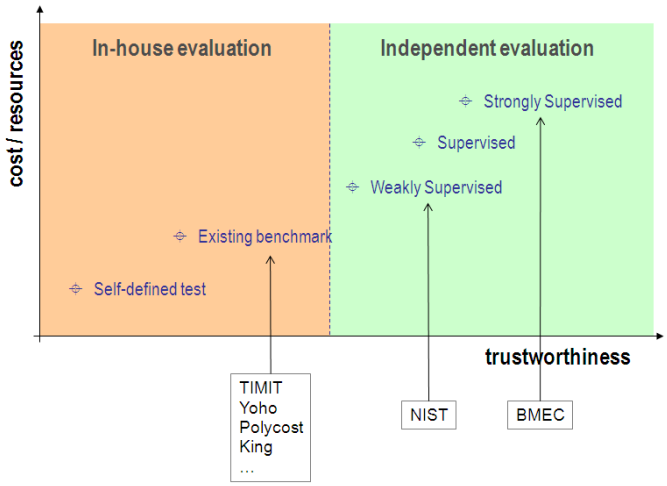

Figure 1: Classification of benchmark evaluations (after (Cappelli et al., 2006).

3.2. Evaluation Methodology

In NIST evaluations, common data-sets, standard measurements of error and evaluation protocol are provided to each participating laboratory. Each evaluation is followed by a workshop so that researchers can compare their submitted results and highlight problems that require further research. For each evaluation, NIST specifies the evaluation tasks and rules in its evaluation plan. The evaluation plan defines the datasets that participants may use during the evaluation procedure. This includes the training data used to train the system, the development data on which participants test and tune their systems and finally the evaluation data to perform the final tests of the system that will be scored by NIST. The evaluation plan includes also the dates for NIST to release the different types of data to the participants, and for the participants to submit their results to NIST.

The submitted results are scored by NIST and the final performances are made available to the participants few weeks after the submission. NIST uses the cost function described above as the basic performance measure (see equation 3). The cost of false rejection C_{FR} has been set as 10 and the cost of false acceptance C_{FA} to 1. The a priori probability of a target P_{tar} has been assigned the value 0.01 and P_{imp} the value 0.99.

3.3. Limitations

NIST SREs have been recognized by the speaker verification scientific community as a key factor for the improvement of the algorithms over the last decade. Nevertheless, they present some limitations:

First, NIST SREs are mostly relevant for applications where the interest is to find if a target speaker is present in a given test speech signal. The mode is fully text independent, i.e. there is no a priori control of what the speaker is saying. Such applications include surveillance applications as well as application for speaker segmentation, clustering or database annotations. Speaker verification has a large potential for commercial applications but in order to be convenient for the users, such systems need to be functional with short training and testing data and with controls to impeach replay attacks. Commercial applications are then mostly based on text-dependent systems and more specifically on text-prompted scenarios, where the speaker is requested to repeat a given utterance. Text-dependent or text-prompted scenarios have never been included in NIST SREs. This fact could explain the lack of commercial vendors participation to those evaluations.

Second, the quality of NIST data may skew the recognition performances of the systems. Indeed speakers could show variabilities due to factors such as topic of conversation, familiarity level with the interlocutor, etc. However the data used by NIST does not control such parameters. For example MIXER corpus was collected in order to support the US government needs with emphasis on forensic-style problem. The main goal was to improve the FBI's Forensic Automatic Speaker Recognition prototype which is designed to be text-independent, channel-independent and to recognize criminals and terrorist talking in different languages (Cieri et al., 2004). For this reason, the focus was more on the languages and channels conditions. This is certainly of interest to the program sponsors but not to the majority of researchers.

Third, the robustness of the systems against impostor attacks is not taken into account by NIST SRE. All impostors access used by NIST are done with zero-effort using so-called random impostures. This means that the impostors attacks are just simulated by testing the target voice against another speaker which is not realistic. Real impostors will of course put more efforts in order to attack the system. This could be by attempting to change their voices, playing a pre-recorded voice or using a text-to-speech system tuned to reproduced voice characteristics close to the one of the target speaker.

Finally, other important aspects of speaker verification systems are also ignored by NIST evaluations, such as the efficiency and the usability. Most of the systems presented in NIST SREs workshops are far away from any real-life commercial context for telephony applications. They require either lots of training data or lots of processing time which is ineffective from the usability point of view.

4. Discussions

There is actually an increasing interest in telephony based speaker recognition applications. Most of the existing commercial applications are text-dependent or text-prompted. So there is an urgent demand to collect relevant databases with which researchers can make a meaningful comparison of different state-of-the-art approaches and assess the progress they could make in this field. Also, and contrary to the NIST SRE data, the acquisition conditions should be as close as possible to real life conditions as encountered in commercial applications. This means short training and testing data, multichannel, mismatched recording conditions, text-prompting, incremental enrollment, etc.

More advanced speaker impostor technologies could also be used such as the impostor voice transformation (Perrot et al., 2005; Matrouf et al., 2006). This technique has been shown to increase the false acceptance rates with the advan-

tage to be low cost in terms of time and human efforts. In the same idea, the detection of replay attacks would also be an interesting challenge oriented towards improving the rejection of impostors. None of these directions have been currently taken by large scale evaluation benchmarks in speaker verification.

Usability tests are also important when considering real-life applications that imply interaction with the user. However, we believe that including such aspects in large scale evaluations are not tractable if it implies an analysis of a life user reaction in front of the system. This is especially true for evaluations such as NIST SRE where there is an increasing number of participants. Nevertheless, one could include more objective criteria oriented towards usability. For example, the duration of the test segments needed by the system to be effective could give an indication on the amount of effort required from the users and would therefore be linked to the acceptance of the system. Also, the system reactiveness could be measured looking at the real time factor of the algorithms. The reactiveness is indeed linked to a good acceptance of the system. Finally, the robustness of the systems in front of longer term intra-speaker variabilities should also be considered because this is the key factor of the user satisfaction.

5. Conclusions

In this paper we presented a review of the current ways to evaluate speaker verification systems, putting an emphasis on the NIST Speaker Recognition Evaluation methods. We tried to put objectively into evidence the current benefits and limitations of such evaluations. NIST SRE is the largest speaker recognition event in which the participating labs can make meaningful comparison of their different approaches with a common evaluation framework and pre-defined protocols. However, the tasks adopted by NIST SRE are, according to us, pretty far away from real-life application and are mostly relevant for applications such as surveillance or mining. Therefore, we believe there is a need of databases and evaluations that are closer to commercial applications of speaker verification systems (short training and testing data, multichannel, mismatched recording conditions, text-prompting, incremental enrollment, etc.). Also, evaluations organizers should include other criteria in order to develop a more user-centered approach. Finally, evaluations should attempt to evaluate stronger forgery scenarios than random impostures as the ability of the system to reject impostor attacks is also an important feature for many applications.

6. References

R. Auckenthaler, M.J. Carey, and H. Llyod-Thomas. 2000. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10.

C. Barras, S. Meignier, and J.-L. Gauvain. 2004. Unsupervised online adaptation for speaker verification over the telephone. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

J.P. Campbell and D. Reynolds. 1999. Corpora for the evaluation of speaker recognition systems. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March.

J.P. Campbell, D. Reynolds, and R. Dunn. 2003. Fusing high- and low-level features for speaker recognition. *In the proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, September.

J.P. Campbell. 1995. Testing with the yoho cd-rom voice verification corpus. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 341–344, May.

R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, and A.K. Jain. 2006. Performance evaluation of fingerprint verification systems. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 28(1):3–18, January.

C. Cieri, J.P. Campbell, H. Nakasone, D. Miller, and K. Walker. 2004. The mixer corpus of multilingual, multichannel speaker recognition data. *In the proceedings of the International Conference on Language Resources and Evaluation*, May.

J. Egan. 1975. *Signal Detection Theory and ROC Analysis*. Academic Press.

A. El Hannani and D. Petrovska-Delacrétaz. 2007. Fusing acoustic, phonetic and data-driven systems for text-independent speaker verificationn. *In the proceedings of Interspeech*, August.

B. Fauve, H. Bredin, W. Karam, F. Verdet, A. Mayoue, G. Chollet, J. Hennebert, R. Lewis, J. Mason, C. Mokbel, and D. Petrovska. 2008. Some results from the biosecure talking face evaluation campaign. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

J. Godfrey, D. Graff, and A. Martin. 1994. Public databases for speaker recognition and verification. *ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, pages 39–42, April.

J. Hennebert, H. Melin, D. Petrovska, and D. Genoud. 2000. Polycost: A telephone-speech database for speaker recognition. *Speech Communication*, 31(2-3):265–270, June.

A.F. Martin and M. Przybocki. 2004. Nist speaker recognition evaluation chronicles. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. 1997. The det curve in assessment of detection task performance. *In the proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 4:1895–1898, September.

D. Matrouf, J.-F. Bonastre, and C. Fredouille. 2006. Effect of voice transformation on impostor acceptance. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May.

H. Melin and J. Lindberg. 1996. Guidelines for experiments on the polycost database. *Proc. COST250 Workshop on The Application of Speaker Recognition Technologies in Telephony*, pages 59–69, November.

J. Pelecanos and S. Sridharan. 2001. Feature warping for

robust speaker verification. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet. 2005. Voice forgery using alisp: Indexation in a client memory. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1, March.

M. Przybocki, A.F. Martin, and A.N. Le. 2006. Nist speaker recognition evaluation chronicles - part 2. *In the proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey)*, June.

D.A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. OLeary, and B. A. Carlson. 1995. The effects of telephone transmission degradations on speaker recognition performance. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May.

D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, January/April/July.

D. Reynolds, W. Andrews, J.P. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, J. Jones, and B. Xiang. 2003. The supersid project: Exploiting high-level information for high-accuracy speaker recognition. *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April.

D.A. Reynolds. 1994. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(3):639–643.

D.A. Reynolds. 1995. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1):91–108, August.

V. Wan and W. Campbell. 2000. Support vector machines for speaker verification and identification. *In proceedings of the IEEE Signal Processing Society Workshop*, 2:775–784.

Year	Corpus	Tasks	Training duration	Testing duration
1996	SWBD I	Speaker detection	2 minutes	3, 10 and 30 seconds
1997	SWBD II p1	Speaker detection	2 minutes	3, 10 and 30 seconds
1998	SWBD II p2	Speaker detection	2 minutes	3, 10 and 30 seconds
1999	SWBD II p3	Speaker detection Speaker tracking	1 minute	30 seconds
2000	SWBD p1 and p2 AHUMADA	Speaker detection Speaker tracking Speaker segmentation	2 minutes	15 to 45 seconds
2001	2000 dataset SWBD I	Speaker detection Speaker tracking Speaker segmentation	2 minutes	15 to 45 seconds
2002	SWBD cellular p1 SWBD p2 and p3 FBI Voice DB	Speaker detection Speaker segmentation	2 minutes, 1, 2, 4, 8 and 16 conversations	15 to 45 seconds and 1 conversation
2003	SWBD cellular p2	Speaker detection	2 minutes, 1, 2, 4, 8 and 16 conversations	15 to 45 seconds and 1 conversation
2004	MIXER	Speaker detection	10, 30 seconds, 5, 15, 40, and 80 minutes	10, 30 seconds and 5 minutes
2005	2004 dataset	Speaker detection	10 seconds, 5, 15, and 40 minutes	10 seconds and 5 minutes
2006	New MIXER data 2005 dataset	Speaker detection	10 seconds, 5, 15, and 40 minutes	10 seconds and 5 minutes

Table 2: History of the NIST Speaker Recognition Evaluations campaigns. The training and testing duration are reported for the speaker detection task only.

Proceedings of the ELRA Workshop on Evaluation Looking into the Future of Evaluation: When automatic metrics meet task-based and performance-based approaches

Edited by Victoria Arranz, Khalid Choukri, Bente Maegaard and Gregor Thurmair

Marrakech, Morocco

27 May 2008

Workshop Chairing Team

Gregor Thurmair, *Linguatec Sprachtechnologien GmbH, Germany* - **chair**
Khalid Choukri, *ELDA - Evaluations and Language resources Distribution Agency, France* – **co-chair**
Bente Maegaard, *CST, University of Copenhagen, Denmark* – **co-chair**

Organising Committee

Victoria Arranz, *ELDA - Evaluations and Language resources Distribution Agency, France*
Khalid Choukri, *ELDA - Evaluations and Language resources Distribution Agency, France*
Christopher Cieri, *LDC - Linguistic Data Consortium, USA*
Eduard Hovy, *Information Sciences Institute of the University of Southern California, USA*
Bente Maegaard, *CST, University of Copenhagen, Denmark*
Keith J. Miller, *The MITRE Corporation, USA*
Satoshi Nakamura, *National Institute of Information and Communications Technology, Japan*
Andrei Popescu-Belis, *IDIAP Research Institute, Switzerland*
Gregor Thurmair, *Linguatec Sprachtechnologien GmbH, Germany*