Computer Speech & Language 63 (2020) 101035

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE

Pavel Matějka*, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, Johan Rohdin, Hossein Zeinali, Ladislav Mošner, Anna Silnova, Ondřej Novotný, Mireia Diez, Jan "Honza" Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czechia

ARTICLE INFO

Article History:
Received 6 May 2019
Revised 23 October 2019
Accepted 15 November 2019
Available online 17 December 2019

Keywords:
Speaker recognition
NIST
Evaluations
GMM
Eigen-channel compensation
JFA
I-vectors
DNN Embedding
X-vectors

ABSTRACT

In this paper, we present a brief history and a "longitudinal study" of all important milestone modelling techniques used in text independent speaker recognition since Brno University of Technology (BUT) first participated in the NIST Speaker Recognition Evaluation (SRE) in 2006—GMM MAP, GMM MAP with eigen-channel adaptation, Joint Factor Analysis, i-vector and DNN embedding (x-vector). To emphasize the historical context, the techniques are evaluated on all NIST SRE sets since 2004 on a time-machine principle, i.e. a system is always trained using all data available up till the year of evaluation. Moreover, as user-contributed audiovisual content dominates nowadays' Internet, we representatively include the Speakers In The Wild (SITW) and VOiCES challenge datasets in the evaluation of our systems. Not only we present a comparison of the modelling techniques, but we also show the effect of sampling frequency.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

In this paper, we want to (i) present the last 13 years of text independent speaker recognition (SR) research and NIST Speaker Recognition Evaluations (SRE)[1] from the perspective of the Brno University of Technology Speech@FIT group[2], (ii) provide some useful "aftermath and lesson-learned" information, and (iii) give a tribute and a thank you to our colleagues from NIST for their series of the NIST SRE's. The paper has two main themes: in the first part (Section 2), we provide a "Tour of speaker recognition at BUT" from the early days until present. It is divided into five main eras, though largely overlapping and sometimes without clear boundaries: GMM era, JFA era, i-Vector era, and two neural network eras, all presented from BUT's stand-point[3] The historical development is put in context with SRE but also with other community events, mainly the seminal Johns Hopkins University (JHU) 2008 workshop and its follow-ups. The paper does not have the ambition to be a full overview of speaker recognition history or a tutorial paper—we recommend excellent tutorials by Campbell (1997), and Kinnunen and Li (2010).

*Corresponding author.
E-mail address: matejkap@fit.vutbr.cz (P. Matějka).

[1] See National Institute of Standard and Technology (0000); Martin and Greenberg (2009, 2010); Greenberg et al. (2013); Sadjadi et al. (2017, 2019) for reference.

[2] Noted as "BUT" or "we" in the rest of the paper.

[3] Compared to a standard research paper, this one thus contains an excessive number of auto-citations.

https://doi.org/10.1016/j.csl.2019.101035
0885-2308/© 2019 Elsevier Ltd. All rights reserved.

The second part of this paper addresses the commonly discussed conflict of "more science" versus "more data". Each edition of SRE was built atop of latest techniques, however, at the same time, more data was available for the development of the evaluation systems. Questions such as "What would have been the performance of neural network (NN) Speaker Recognition (SR) systems on SRE 2006 data, had we had them available back in 2006", or comments such as "just take the simple relevance MAP adapted GMM from the90/s and train it on nowadays huge data, it will probably work the same as your fancy neural nets" are quite frequent. We have addressed these by creating a matrix of five representative techniques (namely, Relevance MAP adaptation, Eigen-channel adaptation, Joint Factor Analysis (JFA), i-vectors with Probabilistic Linear Discriminant Analysis (PLDA) based scoring and neural networks based embeddings (x-vectors) with PLDA scoring) and eight representative data-sets, and performed the analysis. The techniques themselves do not represent the typical SRE submissions (which tend to be fusions of many systems in order to obtain the best results) but rather, they are carefully selected representatives of the state-of-the-art at the given period. The core of this analysis is the standard SRE task: telephone enrollment, telephone test. However, additional experiments are conducted on wideband data (SITW and VOiCES) to analyze a new emerging channel. Our analysis confirms that the NIST SREs brought more and more difficult evaluation data over time, and, as a reaction to these new challenges, more sophisticated and better performing SR techniques emerged. Perhaps interestingly, even at the time of the NIST SRE 2006, a sufficient amount of training data was available to benefit from the newer and more data-hungry techniques such as x-vectors.

The paper is organized as follows: Section 2 presents the history of SRE research at BUT. Sections 3 and 4 contain the details of systems, data and evaluation metrics used for the longitudinal analysis. Section 5 presents its results, and we conclude in Section 6.

2. BUT on SR research and NIST SRE time-line

2.1. End of relevance MAP adaptation era

NIST SRE 2006 was the first speaker evaluation with BUT participation (Matějka et al., 2007a). We were joined by Niko Brummer and Albert Strassheim (then at Spescom Data Voice) and David van Leeuwen (TNO), the submissions therefore carried the label STBU (**S**pescon–**T**NO–**B**UT–**U**niversity of Stellenbosch). At this period, relevance MAP adaptation (Reynolds et al., 2000) of GMM models was the dominating technique in the SR world. We spent significant time on feature analysis, building on our automatic speech recognition (ASR) know-how—we tested RASTA filtering (Hermansky and Morgan, 1994) and heteroscedastic linear discriminant analysis (HLDA) transformation (Kumar, 1997) together with feature mapping (Reynolds, 2003) (simple adaptation done in the feature space aiming at channel-adapted models). More importantly, we implemented eigen-channel adaptation as a simplified variant of JFA (Brummer, 2004; Kenny and Dumouchel, 2004) which made the feature mapping and many of the feature processing tricks obsolete. The superior performance of the eigen-channel adaptation technique in the NIST SRE 2006 attracted a lot of attention in the community and became the new state-of-the-art approach. Eigen-channel adaptation was, in fact, introduced to NIST SRE evaluations already in 2004 by Niko Brummer thanks to his collaboration with Patrick Kenny; it provided excellent performance, but it stayed largely unnoticed, as it was not part of a top-performing system.

In addition to the eigen-channel adapted GMM system, STBU system contained another two approaches based on Support Vector Machines (SVM). The first used GMM supervectors as input to the SVM system and it significantly benefited from Nuisance Attribute Projection (NAP) (Solomonoff et al., 2004), a sub-space based channel compensation technique similar to eigen-channel adaptation. The second approach fed the SVMs with Maximum Likelihood Linear Regression (MLLR) matrices extracted using ASR system (Stolcke et al., 2005). This technique was imported from the ASR field, where MLLR transformations were typically used for speaker adaptation. The SVM-MLLR system worked the worst, but it fused well with the other systems. Niko Brummer also contributed with proper normalization (t-norm) (Auckenthaler et al., 2000) and system fusion techniques (e.g. Logistic Regression) using his then recently introduced FoCal toolkit[4], which remains an invariable part of our systems till this day.

Two papers in 2007 Special issue of IEEE T-ASLP give full account of our work for SRE 2006: Brümmer et al. (2007) provides full system description including SVM sub-systems and (Burget et al., 2007) concentrate on the analysis of feature extraction and channel compensation in the GMM system.

2.2. JFA era

JFA as a technique for subspace modeling of speaker and channel variabilities in the space of GMM parameters was introduced by Patrick Kenny already in 2003 (Kenny et al., 2003). However, it was only after the success of eigen-channel adaptation that the community became interested in JFA. This interest was boosted by the superior results obtained with the full JFA model as first presented on NIST SRE data by Kenny et al. (2008), and later confirmed in NIST SRE 2008 evaluation by several sites (Kajarekar et al., 2009; Sturim et al., 2009; Burget et al., 2009a).

SRE 2008 witnessed the top of JFA era—BUT submission included two full JFA systems combined with one SVM-MLLR, similarly as in the previous evaluations. Significant amount of work done for SRE 2008 did not make it to the final submission, but generated interesting research ideas: use of phonotactics in SR (where an attempt was made to piggy-back on our excellent results in LR) and the first attempts to use prosody and cepstral contours, later elaborated in the PhD thesis of Kockmann (2012)

[4] https://sites.google.com/site/nikobrummer/focal

and his publications. Interspeech 2009 paper (Burget et al., 2009a) contained description and analysis of the whole BUT submission, and Burget et al. (2009b) presented at the same venue thoroughly analyzed the individual variants and simplifications of JFA. It pointed out the importance of score normalization (ZT-norm) for full JFA and the advantage one could gain by training additional eigen-channels on (although limited) microphone data to improve robustness to channel variations.

Johns Hopkins University 2008 summer workshop was a key moment in the JFA era[5]. The work included conditioning of JFA, fast techniques for JFA scoring (later summarized in Ondrej Glembek's thesis Glembek, (2012) and served as a basis for fast i-vector scoring) and early work on discriminative variants of JFA. At the workshop, Najim Dehak also started elaborating on the idea of using JFA for extracting speaker factors as low-dimensional fix-length vectors encoding the speaker identity in an utterance, which eventually led to the introduction of i-vectors (Dehak et al., 2011) (see next section). In addition to the technical achievements, this workshop laid grounds to intensive community work in SR that persists until nowadays.

2.3. I-vector era

As mentioned above, the 2008 workshop devoted significant efforts to the study of JFA and different ways of producing verification scores with the model. Large amount of follow-up work concentrated on using JFA to extract fixed-length utterance representations as *features* for another classifier. Najim Dehak experimented with the channel factors (where any speaker-related information should be suppressed) and found them to be still good features to characterize speakers. This led him to the definition of total variability vectors as the sole low-dimensional representations of utterances. The paper usually cited is the journal one (Dehak et al., 2011) but although it was submitted in 2009, it took almost two years to get published. The first publication on total variability vectors (still not called i-vectors) is Dehak et al. (2009), and, as usual, the word on the superior performance of these vectors was spreading fast in the community. The term i-vector was coined around 2009/2010, where the "i" stood for "identity" or "intermediate".

NIST SRE took place early in 2010 and we formed the ABC consortium of **A**gnitio South Africa (Niko Brummer), **B**UT and **C**RIM Canada (Patrick Kenny) Brummer et al. (2010). That year, NIST set new parameters for the decision cost function (DCF) so that we had to inflate the number of non-target trials in our development set, but the main focus was on i-vectors. Probabilistic linear discriminant analysis (PLDA) (Ioffe, 2006; Prince, 2007) was shown to be an excellent scoring tool for the i-vector framework and Patrick Kenny has compared Gaussian PLDA with heavy-tailed (HT) PLDA, concluding that HT-PLDA worked even without score normalization, better than Gaussian PLDA with score normalization; the advantage was however unclear with non-matching channels Matějka et al. (2011b). The era of traditional JFA used directly for scoring was over.

Summer 2010 saw Brno as the world's center of SR: we hosted the SRE 2010 workshop, immediately followed by Odyssey workshop, with notable Patrick Kenny's invited talk on HT-PLDA Kenny (2010). Both events confirmed i-vectors as the ruling SR paradigm. Immediately following was the 5-week BOSARIS workshop[6] coordinated by Lukas Burget, Patrick Kenny, and Niko Brummer. The workshop generated a significant amount of work and numerous papers that were influencing our SR activities in the following years, especially Ondrej Glembek's work on the efficient implementation of i-vector extraction (Glembek et al., 2011b). Both i-vector extraction and PLDA were generative models, therefore, the first attempts were made to train them discriminatively (Burget et al., 2011; Glembek et al., 2011a; Cumani et al., 2011). The workshop also produced Niko Brummer's BOSARIS toolkit for calibrating, fusing and evaluating SR scores[7] that was and still is widely used in the community.

I-vectors dominated our work in SR for several following years. Several important PhD theses document this period and we recommend them as reference reading for anyone wishing to acquire broader context than from conference papers: Glembek (2012) summarized scoring techniques for both JFA based models and i-vectors, with an accent on efficient implementation, Plchot (2014) departed from discriminative training of PLDA and was among the first to take into account the uncertainty in i-vector estimates to obtain more reliable PLDA scoring and Kockmann (2012) re-defined the continuous i-vector model to subspace multinomial model (he was using categorical prosody feature) that has reached other domains including topic detection (Kesiraju et al., 2016) and language modeling (Beneš et al., 2018).

During 2009−11, we cooperated with SRI STAR Laboratory in the IARPA-funded BEST program: in addition to work on high-level features for SR, we have defined and made available the PRISM evaluation set (Ferrer et al., 2011) that has been serving to evaluate variabilities in language, channel, speech style and vocal effort, including new types not available at the time such as severe noise, and reverberation. We participated in SRE 2012 but we do not describe it here, as it was different to all other evaluations (see Section 4.2). The DARPA-sponsored RATS project was fully occupying us, including its own evaluations. As we had to work with very noisy and degraded channels, we spent efforts on noise-robust features (Plchot et al., 2013), robust VAD (Ng et al., 2012) and we experimented with SVM-based fusion using side information (Plchot et al., 2013).

The multi-week workshops gained popularity in the community; JHU 2008 and BOSARIS 2010 were followed by another BOSARIS in 2012[8], and another JHU workshop in 2013[9] combining SR and language recognition (LR), as both are sharing a number of techniques. JHU 2013 brought up the Domain Adaptation Challenge 2013 (nick-named as the "Doug's

[5] work-group "Robust Speaker Recognition Over Varying Channels led by Lukas Burget, https://www.clsp.jhu.edu/workshops/08-workshop/robust-speaker-recognition-over-varying-channels/ links also to the final report.
[6] Brno Speaker Recognition Summer Workshop, https://speech.fit.vutbr.cz/workshops/bosaris2010
[7] https://sites.google.com/site/bosaristoolkit/
[8] https://speech.fit.vutbr.cz/cs/short-news/second-bosaris-workshop-2012-started-14-people-arrived
[9] https://www.clsp.jhu.edu/workshops/13-workshop/speaker-and-language-recognition/

challenge") (Garcia-Romero and McCree, 2014; Villalba and Lleida, 2014), where different ways of adapting the PLDA model were mainly studied (Glembek et al., 2014). Another two workshops took place in Torino, Italy (TOSREW, 2015[10]) and Stellenbosch, South Africa (ASRWIS, 2017[11]), which were already investigating into NN techniques in SR.

2.4. Neural network era

2.4.1. Hybrid i-vector/neural network era

Our first successful attempts to use NNs for SR used the i-vector/PLDA pipeline, by integrating pre-trained NNs in different ways to improve the SR performance. In the first approach, pioneered by SRI Lei et al. (2014), a DNN pre-trained for senone classification (i.e. ASR acoustic model) was used to align speech frames to Gaussian components in the i-vector extractor model (i.e. the Gaussian components are forced to correspond to phonetic classes). This approach led to more robust i-vector estimation and improved SR performance. The second approach introduced NN-based bottleneck features (BNF) (Grézl et al., 2007) instead of (or together with) MFCC features. Again, the bottleneck features were optimized to discriminate phonemes in contrast to architectures directly trained for speaker discrimination described in the next section. We performed numerous experiments with both approaches (Novotný et al., 2016; Lozano et al., 2016). To address the language-dependency issue of BNFs, we used multilingual training of BNFs introduced by Karel Vesely in 2012 for the IARPA Babel program (Veselý et al., 2012) and built on excellent results of these features in LR (Fér et al., 2015). The conclusion of (Matějka et al., 2016) was that concatenated BNFs and MFCCs indeed worked the best, and there was no need for an additional DNN-based alignment.

The year 2016 was a true evaluation year: the first Speakers in the Wild challenge (organized by SRI) took place in winter[12] and brought signals from difficult environments, with reverberation, noise and low quality recording devices. We obtained excellent results with a system based on both the spectral features as well as on the BNFs (Novotný et al., a). SITW, for the first time, showed the importance of diarization (we used our Bayesian HMM approach with eigen-voice priors (Diez et al., 2018)). An important part of system development was data *augmentation*, a step that has become inevitable in all our following work.

Late summer and fall were dominated by SRE 2016; we formed the "ABC" consortium consisting of **A**gnitio, **B**UT and **C**RIM. 2016 was the last year with only i-vector systems in the final submission (Plchot et al., 2019). A bitter surprise came with the inferior performance of BNFs, for which the language variability (different languages in development and evaluation sets) was the most obvious reason. Discriminative PLDA was included in the final system; while it did not provide superior performance itself, it fused well with the other techniques. Score normalization was (again) an issue: an analysis in Matějka et al. (2017) has shown that adaptive s-norm—with the careful cross-language and cross-channel selection of speaker cohorts—worked the best.

2.4.2. Era of neural network embedding and end-to-end systems

Our NN efforts continued with the development of fully end-to-end architecture where building blocks of a standard i-vector/PLDA system were gradually replaced by corresponding NNs (Rohdin et al., 2018). We showed that it is possible to benefit from end-to-end training if we constrain the system not to deviate too much from a standard i-vector+PLDA one: it consists of a DNN module for extraction of sufficient statistics (**f2s**), a DNN module for extraction of i-vectors (**s2i**) and finally, a discriminative PLDA (DPLDA) model (Burget et al., 2011; Cumani et al., 2013) for producing scores. These three modules are first developed and trained individually so that they mimic the corresponding part of the i-vector+PLDA baseline. After that, they are combined and the system is further trained in an end-to-end manner on both long and short utterances. In contrast to the previous section, the whole architecture is now trained for the final task of speaker discrimination.

At the same time, we followed the efforts of using DNN embeddings in SR and were happy that our colleagues at JHU succeeded in making the system work: first on short utterances with the abundance of training data (Snyder et al., 2016) (where the embeddings were extracted from a DNN trained with speaker verification objective, i.e., binary task of pair-wise speaker comparison) and finally on standard SR tasks (Snyder et al., 2018) (where the training follows a speaker identification criterion). Like many others, we appreciated the Kaldi recipe, that we analyzed, for example, for performance in noise, with various NN architectures and with data augmentation (Novotný et al., 2018b).

Our participation in SRE 2018 (Alam et al., 2018) fell fully into the NN era. We worked again in the "ABC" consortium with Nuance (Niko Brummer), CRIM, Omilia and UAM Madrid. X-vectors were clearly dominating and the pre-processing of data (augmentation and chunking into segments) were found to be crucial for good system performance. Hossein Zeinali produced a useful TensorFlow implementation of the x-vector extractor training[13] complementing the Kaldi recipe. Encouraging results were obtained with revived HT-PLDA (Kenny, 2010).

Our current work aims at several directions, with some of the results yet to be published: in far-field SR, we use classical and neural microphone array pre-processing, data augmentation and system (usually PLDA) re-training (Mošner et al., 2018), we experiment with adversarial adaptation applied to x-vectors (Rohdin et al., 2019) and attention modeling added to x-vector extraction NN, and (in good tradition of the BUT group) into interactions and cross-domain issues with LR, diarization, ASR, VAD, and others.

[10] https://areeweb.polito.it/SRG/tosrew2015/
[11] https://agnitiosa.github.io/ASRWIS/
[12] http://www.speech.sri.com/projects/sitw/
[13] https://github.com/hsn-zeinali/x-vector-kaldi-tf

3. Speaker modeling techniques for text-independent speaker recognition

In this section, we briefly describe the main modelling techniques used in the NIST SRE which are used in the longitudinal study in Section 5. Some of the techniques use primarily unlabeled data (GMM, i-vector), however, the main driver for improvement and evolution of SR algorithms is the availability and more efficient use of ever bigger amounts of labeled training data. Without huge quantities of labeled data, it would not be possible to incorporate powerful, but data hungry techniques (PLDA, x-vector). In addition to the main techniques, we briefly describe two *auxiliary techniques*, namely score normalization and data augmentation. In the experiments, we use these techniques only for the main techniques that need them (see Sections 3.8 and 3.9). The model parameters and settings used in the experiments are given in Table 6.

3.1. Gaussian mixture models

Gaussian mixture models (GMMs) have been an integral part in speaker recognition for more than two decades. Their main assumption is that the features of an utterance are generated from a Gaussian mixture model whose means depend on the *speaker's identity* as well as on *channel effects*. The concatenated vectors of parameters (typically means) are usually referred to as a *supervector*. In order to do speaker recognition, we need to estimate the supervector of means from the features and also, ideally, remove the effect of channels from it. In most speaker recognition scenarios, there is not enough enrollment (or test) data available to reliably estimate the supervector of means using maximum likelihood (ML). Instead, we assume a prior on the supervector and use its maximum a posteriori estimate. To mitigate the effect of channels, the supervector, $\mathbf{M}$, is further decomposed into a global mean, $\mathbf{m}$, a speaker-specific part, $\boldsymbol{s}$, and a channel (utterance) specific part, $\boldsymbol{c}$, i.e.,

$$\mathbf{M} = \mathbf{m} + \mathbf{s} + \mathbf{c}. \tag{1}$$

This approach has been common to all dominant techniques in the NIST SRE from 1996 to 2016. The difference lies in the choice of speaker and channel components. The most common choices are summarized in Table 1 and described briefly in the following sections. All the approaches involve a *universal background model* (UBM) which is a GMM estimated on features from many utterances of many speakers. The supervector of the UBM means serves as $\mathbf{m}$ in Eq. (1). Further on, let $\boldsymbol{\Sigma}$ be a (block) diagonal matrix with the covariance matrices of the individual Gaussians on the diagonal. In our experiments, we use GMMs with 2048 Gaussians.

3.2. Relevance-MAP

Relevance MAP was originally proposed for speaker adaptation in automatic speech recognition in Gauvain and Lee (1994) and adopted for speaker recognition in Reynolds et al. (2000). This approach assumes that the speaker specific component of the mean supervector, $\mathbf{M}$, is given by $\mathbf{s} = \frac{1}{\sqrt{\tau}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}$, where τ is the so-called *relevance factor* and $\mathbf{z}$ is a latent *speaker variable* of the same dimensionality as the supervector with standard normal prior. This is equivalent to saying the speaker-specific component of the mean supervector $\mathbf{s}$ has prior $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \frac{1}{\tau}\boldsymbol{\Sigma})$. We refer to the method as *GMM*. Scoring is done by obtaining a point MAP estimate of $\mathbf{M}$ for the enrollment utterance and then calculating the likelihood of the test data given this model versus the likelihood of the test data given the UBM (feature LLR). Note that with this model, we do only short time mean and variance normalization on the feature level to deal with channel effects. In our experiments, we use $\tau = 19$ (Burget et al., 2007).

3.3. GMM eigen-channel compensation

The GMM eigen-channel model (GMM-EC) (Kenny et al., 2003; Burget et al., 2007) extends the GMM approach by adding a channel component $\mathbf{c} = \mathbf{U}\mathbf{x}$ (see Table 1), where $\mathbf{x}$ is a *channel variable* with standard normal prior. The matrix $\mathbf{U}$ is estimated by PCA on supervectors from which the speaker specific mean supervector has been subtracted. Scoring can be done in many ways. In this paper, we use linear scoring (Glembek et al., 2009). We set the number of columns in $\mathbf{U}$ (i.e., the number of eigen-channels) to 50 (Burget et al., 2007).

Table 1
Speaker and channel component of the supervector of means assumed by different GMM-based models. The matrices $\mathbf{V}$, $\mathbf{U}$ and $\mathbf{T}$ have many fewer columns than rows. The matrix $\mathbf{D}$ is diagonal. The *factors*, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{w}$, follow standard normal distribution. Training and scoring refer to the most common approaches which are also used in this paper. EM refers to Expectation Maximization and MD to Minimum divergence. Note that PLDA also models a speaker and a channel component but in the i-vector space.

Method	Speaker, s	Channel, c	Training	Scoring
GMM	$\frac{1}{\sqrt{\tau}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}$	-	-	Feature LLR
GMM-EC	$\frac{1}{\sqrt{\tau}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}$	$\mathbf{U}\mathbf{x}$	PCA	Linear
JFA	$\mathbf{V}\mathbf{y}+\mathbf{D}\mathbf{z}$	$\mathbf{U}\mathbf{x}$	EM+MD	Linear
i-vector	$\mathbf{T}\mathbf{w}$		EM+MD	PLDA

Table 2
x-vector topology proposed in Snyder et al. (2019). K in the first layer is used to indicate using different features with different dimensions and N is the number of speakers.

Layer	Layer context	**(Input)** $\times$ **output**
frame1	$[t-2, t-1, t, t+1, t+2]$	$(5 \times K) \times 512$
frame2	$[t]$	512×512
frame3	$[t-2, t, t+2]$	$(3 \times 512) \times 512$
frame4	$[t]$	512×512
frame5	$[t-3, t, t+3]$	$(3 \times 512) \times 512$
frame6	$[t]$	512×512
frame7	$[t-4, t, t+4]$	$(3 \times 512 \times 512$
frame8	$[t]$	512×512
frame9	$[t]$	512×1500
stats pooling	$[0, T)$	1500×3000
segment1	0	3000×512
segment2	0	512×512
softmax	0	$512 \times N$

3.4. Joint factor analysis

The Joint factor analysis (JFA) model extends the GMM-EC model by replacing its (rather ad hoc) speaker component with $\mathbf{s} = \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}$ where $\mathbf{y}$ and $\mathbf{z}$ are standard normal distributed latent *speaker variables*, the matrices $\mathbf{V}$ and $\mathbf{D}$ are learned from data. For learning $\mathbf{V}$, $\mathbf{D}$ and $\mathbf{U}$, we use the EM algorithm (Dempster et al., 1977) and minimum divergence. As for GMM-EC, we use linear scoring. We set the number of columns in $\mathbf{V}$ (the number of eigen-voices) to 300, and $\mathbf{U}$ (the number of eigen-channels) to 100 (Burget et al., 2009b). The matrix $\mathbf{D}$ is diagonal.

3.5. i-vectors

The i-vector (Dehak et al., 2010) model simplifies the previous models by replacing the speaker and channel component with one component that is assumed to model both speaker and channel variability. The model is

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{2}$$

where the matrix $\mathbf{T}$ is a model parameter (often referred to as an *i-vector extractor*) and $\mathbf{w}$ is a latent variable with standard normal distribution. The latent variable, $\mathbf{w}$, is specific to the utterance but, contrary to the channel component of the previous models, the i-vector extractor is trained in such a way that $\mathbf{T}\mathbf{w}$ captures both the speaker and the channel contribution to the supervector of means. The MAP estimate of $\mathbf{w}$ is known as the *i-vector*. For scoring, the i-vectors are computed for both the enrollment and the test utterance and then a backend model is used for removing channel effects and comparing the two i-vectors. Typically, PLDA (see Section 3.7) is used as a backend model. In our experiments, we set the number of columns in $\mathbf{T}$ (i.e., the i-vector dimensionality) to 600 (Matějka et al., 2011b).

3.6. Neural network embedding: x-vector topology

Current state-of-the-art systems for text-independent speaker verification are based on NN embeddings. Similarly to i-vectors, these embeddings are low dimensional fix-length representations of utterances, which are however obtained using NNs discriminatively trained to extract only speaker-specific information. Different NN architectures and training objectives were proposed for this purpose (Snyder et al., 2017). In our experiments, we use the first truly successful and currently the most popular architecture for extracting the so-called x-vector (Snyder et al., 2018).

The NN for x-vector extraction is composed of several Time-Delay NN (TDNN) layers[14] which operate in a frame-by-frame manner. The TDNN layers are followed by a global pooling layer. This layer estimates the mean and the standard deviations of the outputs of the last TDNN layer over time in order to obtain the fixed-length utterance representation. Additional layers are used to reduce the dimensionality of such representations to obtain the final embedding/x-vector. For training, one more softmax layer is added, which serves as the classifier of training speaker identities. The exact architecture used in our experiments is summarized in Table 2. For more details on robustly training the x-vector architecture see the original work (Snyder et al., 2018). For training the x-vector, we used original Kaldi recipe[15] where the system is trained on original speech segments together with their augmentations.

[14] TDNN layer performs a one-dimensional convolution over time with dilation of more than one frame as used in Waibel et al. (1989) and Peddinti et al. (2015) for speech recognition.

[15] https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

3.7. Backend—PLDA

To facilitate comparison between i-vectors or x-vectors and obtain a speaker verification score, Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe, 2006; Prince, 2007; Kenny, 2010) is often utilized. There are many variants of PLDA, the most common of which—and the one used in this work—assumes that the distribution of i-vectors or x-vectors, ϕ, is modeled as

$$\boldsymbol{\phi} = \overline{\boldsymbol{\phi}} + \mathbf{V}\mathbf{y} + \boldsymbol{e}, \tag{3}$$

where, ϕ is an observed vector (i-vector or x-vector), $\overline{\boldsymbol{\phi}}$ is a global mean of the observed data, $\mathbf{y}$ is a latent *speaker variable* with standard normal prior and $\boldsymbol{e}$ is a latent *channel variable* with prior $p(\boldsymbol{e}) = \mathcal{N}(\boldsymbol{0}, \mathbf{W})$. Note that this model is conceptually very similar to JFA. Typically, and in this work, the parameters of the model $\mathbf{V}$, $\mathbf{W}$ are estimated by the EM algorithm (Ioffe, 2006; Prince, 2007).

Usually, i-vectors are subjected to several pre-processing steps before being modeled by PLDA. In this paper, we use Linear Discriminant Analysis—reducing the i-vector dimension to 250—followed by length-normalization (Garcia-Romero and Espy-Wilson, 2011). It is worth noting that generalizing the Gaussian assumption to more heavy-tailed distributions (Kenny, 2010) can reduce the need for such preprocessing steps. However, the Gaussian assumption leads to a closed-form solution for the LLR used as a verification score.

3.8. Score normalization

The goal of score normalization is to reduce within-trial variability leading to improved performance, better calibration, and a more reliable threshold setting (Matějka et al., 2017). In our experiments, the effect of score normalization was large for JFA but negligible for the other systems. Therefore we report the results with score normalization (ZT-norm) only for systems based on JFA.

3.9. Data augmentation

As mentioned in Section 3.6, the Kaldi recipe applies data augmentation in x-vector training. Each utterance is augmented with approximately two new versions that are corrupted by either noise, reverberation, music or babble. Augmenting the training data has been shown to greatly improve the training of x-vector extractors while having a minor effect on the training of i-vector extractors (Snyder et al., 2018; Novotný et al., 2018b). Since the older methods are either structurally similar to i-vectors (GMM-EC, JFA) or much less data-hungry (GMM), we assume data augmentation would also not be beneficial for these methods. Therefore, we do not use data augmentation for them.

4. Data and evaluation metric

4.1. Training data

Training data for NIST SRE experiments are formed by the data released by NIST before the year of that evaluation campaign plus data from Switchboard. We performed also a comparison of systems trained on Voxceleb only.

4.1.1. Switchboard

NIST SRE data between the years 1999 and 2003 were derived from Switchboard data collections targeted for the speaker recognition where each speaker had to make several calls. All speakers are English speaking. We used several collections from Switchboard:

- Switchboard 2 Phase I (Graff et al., 1998), II (Graff et al., 1999) and III (Graff et al., 2002) released in 1998, 1999 and 2002 with 657, 679 and 640 participants, respectively. It contains only land-line calls.
- Switchboard Cellular Part I (Graff et al., 2001) and II (Graff et al., 2004) released in 2001 and 2004 with 254 and 419 participants, respectively. It focuses mainly on cellular phone technology under varied environmental conditions.

4.1.2. NIST speaker recognition evaluations

Training data from NIST SRE are formed by all training data from a particular evaluation year. Overall data statistics from individual SREs are listed in Table 3 and more information about each NIST dataset can be found in Section 4.2.

4.1.3. VoxCeleb

VoxCeleb is a recently introduced (2017) audio-visual dataset consisting of short clips of speech, extracted from celebrity interview videos uploaded to YouTube. The dataset consists of two parts, VoxCeleb1 and VoxCeleb2 (Nagrani et al., 2017). Each part has its own train/test split and there is no overlap of speakers between the two parts. There is however a small overlap with

Table 3
Training data statistics. Note that for x-vector training, the data is augmented with approximately two corrupted versions of each utterance.

Dataset	#speakers	#files/speaker	net speech [h]
SWB	2594	10.9	1139
SRE04	310	26.4	197
SRE06	2228	15.0	946
SRE08	1328	20.2	724
SRE10	506	15.4	286
SRE16	221	53.5	121
SWB + SRE04	2904	12.5	1336
SWB + SRE04−06	5217	17.9	2823
SWB + SRE04−08	6413	18.7	3547
SWB + SRE04−10	6919	20.6	3996
SWB + SRE04−16	7140	22.1	4138
VOXCELEB	7146	23.2	2420

SITW dataset which was mitigated by removing the affected speakers from Voxceleb. In total, there are 166 thousand audio files (distributed in 1.2 million speech segments) spanning 7146 speakers.

4.2. Evaluation data

The experiments are evaluated on the *most relevant*[16] *core conditions* of the NIST SRE datasets from 2004 to 2018. Furthermore, to provide the reader with a more complete analysis of results, the challenging Speakers In The Wild (SITW) Mitchell McLaren (2016) and the recent VOiCES datasets (Nandwana et al., 2019) are also considered as benchmarks. The main characteristics of these datasets are described below.

NIST SRE datasets evolved through the years: in SRE04 (1side-1side) enrollment utterances consist of 5-minute conversational excerpts collected over telephone channels. Utterances contain mainly English speech, but can also contain Arabic, Mandarin, Russian and Spanish. The test utterances have a similar length as the enrollment but could be collected using non-telephone channels, and are rarely non-English.

The SRE06 (1conv4w-1conv4w) is very similar to SRE04, although test utterances are now limited to only telephone channels (as the enrollment). On the other hand, the variability in languages is extended to the test.

In SRE08 (short2-short3—det6 telephone-telephone) the enrollment and test utterances consist of telephone conversational excerpts of around 5 minutes which can sometimes contain "other languages" than English.

For SRE10 (condition 5—telephone-telephone), all data is limited to English speech. The characteristics of enrollment and test utterances are similar to those from SRE08.

For SRE16, we consider the Cantonese evaluation data[17]. It consists of family telephone conversations in which speakers were encouraged to use different telephones for each session.

SRE18 was similar to SRE16 but with telephone speech solely in Tunisian Arabic. SRE18 contains also a preview of the wideband *audio from video* data (VAST collection), which follows a different protocol for evaluation when multiple speakers can be present in the test segment. Our systems for VAST data were developed separately and included speaker diarization. In our analysis, we focus only on the telephone part of the SRE18—the Call My Net 2 collection (cmn2).

The SITW core-core eval dataset (Mitchell McLaren, 2016) is recorded in 16kHz (these data were downsampled for the 8kHz experiments). It contains English speech recorded in non-controlled situations, which results in different levels of degradation, noise types and different microphones for each of the sessions. The enrollment and test utterances are continuous excerpts containing speech of a single speaker, and the test utterances have a variable length between 6 and 180 s.

Finally, VOiCES eval set (Nandwana et al., 2019) is also recorded in 16 kHz, and contains clean read English speech retransmitted in rooms of different sizes, which have different room acoustic profiles. The recordings were made with different background noises played concurrently.

Table 4 provides statistics of these datasets regarding the number of files, speakers, amount of speech and trials.

4.3. Evaluation metrics

We report all results in terms of Equal Error Rate (EER), which is a common measure characterizing the performance of the biometric system. It is a well-known operating point, where the false alarm rate and miss rate are equal. It can be shown that this

[16] Given the high amount of overlap between NIST SRE 2005 and NIST SRE 2006, only the latter dataset was considered. The NIST SRE 2012 dataset was also not used in this study, as in 2012 the evaluation was significantly different from other years: prior knowledge of some target speakers was allowed for computing the trial detection scores and a different evaluation metric was used.

[17] This is a subset of the real evaluation set

Table 4
Evaluation data statistics. Note that the "speech" column indicates the net speech.

Dataset & Condition	#files	#spk	speech [s]	#trials	
	enroll/test		enroll/test	tgt	nontgt
SRE04 1side-1side	616/1174	310	143/140	2.38k	23.8k
SRE06 1conv4w-1conv4w all	808/2668	598	134/135	3.62k	47.5k
SRE08 short2-short3 det6	1788/2569	1033	137/138	2.68k	33.2k
SRE10 condition 5, extended	4267/767	438	133/141	7.17k	409k
SRE16 yue	601/4858	100	60/35	193k	946k
SRE18 cmn2	1316/12135	290	61/35	60.7k	2.0M
SITW core-core	1204/1204	299	31/31	3.66k	718k
VOiCES eval	328/11069	300	12/13	36.4k	3.6M

point acts as a scalar summary of the whole system (or the DET curve—Detection Error Trade-off curve (Martin et al., 1997)) and it is insensitive to calibration. The value of EER gives a rough idea, how close the DET curve is to the origin and therefore, its value can serve as an approximate comparison between the systems. Even though the properties of this measure seem to be attractive, it is not very useful in practical applications which usually operate either in a region of low false alarm rate (e.g., authentication systems) or low miss rate (e.g., law enforcement).

To compare systems at an interesting operating point, NIST has defined the Decision Cost Function (DCF). With some parameter adjustments or modifications in its definition, it has served as the primary criterion in every NIST SREs. It is designed to consider the overall cost of making the two types of detection errors (miss and false alarm).

However, since the DCF parameters are set differently for individual evaluations and we only want to compare general discriminability of the systems and techniques across different NIST test sets, we will use EER in all our experiments.

5. Longitudinal analysis

In this section, we will evaluate all methods on individual NIST benchmarks as well as on SITW and VOiCES challenges. We will concentrate on individual NIST SREs by making a little trip to history and training each method as if it was the time of each particular evaluation. We will also look at the effect of different sizes of training data on the performance of each method. Using the out-of-domain VOXCELEB dataset, we will perform an analysis of the difficulty of particular NIST SREs and finally, we will offer a sneak-peek into relatively more challenging wideband data obtained mostly from YouTube videos and other networks (SITW) or wideband data corrupted by reverberation and noise (VOiCES).

5.1. Redoing NIST SREs with all major techniques

In Fig. 1, we can compare the performance of different techniques (GMM, GMM-EC, ivec, xvec) as if they were run at the time of individual NIST SREs. Important details about the size of each model and its input feature configuration can be found in Table 6.

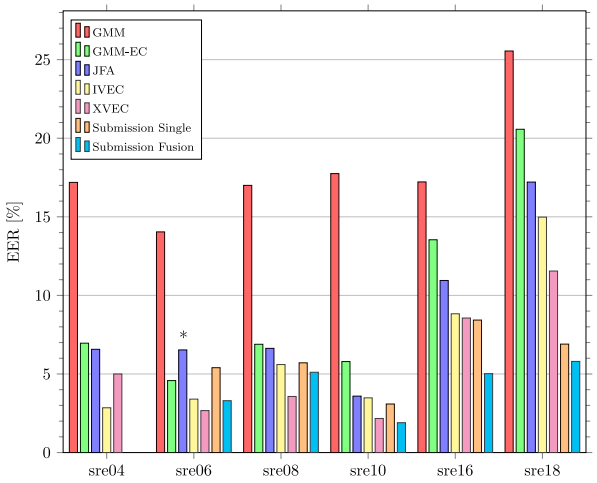

Fig. 1. Equal error rates of all techniques as if they were run for particular NIST SRE. The last two columns starting in SRE06 represent our best single system and a fusion of our submissions to the particular evaluation. Note the out-of-trend JFA EER for SRE06 condition (marked with a * symbol): for sake of consistency with all other experiments, we used the whole training set for ZT-norm. If, however, we exclude the SWB set from ZT-norm, the EER drops significantly to the in-trend region.

Table 5
BUT systems submitted to particular NIST SREs (single best system and primary fusion).

Type	Year	Ref.	System Description
single	2004		-
	2006	Matějka et al. (2007a)	MFCC-HLDA+GMM+Eigen-channel compensation
	2008	Burget et al. (2009a)	MFCC, Gender Dependent JFA + ZT-norm
	2010	Matějka et al. (2011b)	MFCC FullCov ivec, PLDA
	2016	Plchot et al. (2019)	MFCC FullCov ivec, DPLDA
	2018	Alam et al. (2018)	x-vectors + PLDA
fusion	2004		-
	2006	Matějka et al. (2007a)	11 systems (GMM-EC, GMM-SVM, MLLR-SVM,...)
	2008	Burget et al. (2009a)	3 systems (2xJFA+MLLR-SVM)
	2010	Matějka et al. (2011b)	8 systems (JFA + ivec + MLLR-SVM,...)
	2016	Plchot et al. (2019)	fusions of 7xBUT+3xAGN+8xCRIM
	2018	Alam et al. (2018)	3 x-vectors systems

Table 6
Configuration of systems representing individual major SR techniques. *MFCC_E_D_A* denotes MFCC+Energy+Delta+DoubleDelta, *ShortCMVN* refers to short time mean and variance normalization over 301 frames, *F-b-F* refers to frame-by-frame scoring.

Name	Features	Model	Scoring
GMM	19 MFCC_E_D_A ShortCMVN	GMM(2048)	F-b-F
GMM-EC	19 MFCC_E_D_A ShortCMVN	GMM(2048) Eigen-channel comp(50)	linear scoring
JFA	19 MFCC_E_D_A ShortCMVN	GMM(2048) U(100), V(300), Z	ZT-norm linear scoring
IVEC	19 MFCC_E_D_A ShortCMVN	GMM(2048) - IVEC(600) L2norm - LDA(250)	PLDA
XVEC	20 MFCC ShortCMN	DNN (topology in Table 2)	PLDA

Since SRE06, we also include a single best system and a primary fusion of our (BUT + consortium) submission to the particular NIST SRE. In Table 5, we can also observe how our single-best system changed and what combination worked as the best fusion in each evaluation since SRE06. Table 5 also contains references to corresponding system descriptions of our NIST submissions. As expected, overall results in Fig. 1 show that a newer technology provides better results.

Starting with SRE04, where we used only Switchboard dataset for training, we can observe relatively poor performance of x-vectors, especially compared to i-vectors but also compared to JFA and GMM-EC. There is probably simply not enough training data for the model of this size and design. Another possible reason is the fact that Switchboard contains only English calls, while the SRE04 test set comes from 5 languages. This might, apart from the lack of training data, cause the discriminatively trained x-vector not to generalize well for these unseen conditions.

In NIST SRE 2006, we already start to observe expected trends of newer techniques outperforming older ones, or in the case of JFA matching the performance of GMM-EC (with a ZT-norm cohort without Switchboard). We can also observe that the result of the single best system submitted to SRE06 is worse than the same technology (GMM-EC) trained for the purposes of this paper. The reason for this is that in 2006 we trained the system only on data from NIST SRE 2004 and 2005, while now we train also with Switchboard. We can also see that x-vectors alone are already better than our primary fusion back in 2006.

NIST SRE 2008 came with a harder test set and more training data available to the community. This helped the x-vector system to outperform i-vectors more significantly. Our single-best system in SRE08 was based on JFA and its performance is roughly the same as our i-vector based system (IVEC). The reason for a slightly better performance of JFA in our 2008 submission may be the fact that it is a gender-dependent system with gender-dependent ZT-norm. Again, today's state-of-the-art x-vectors are better than our final fusion in 2008. Interestingly enough (at least for the telephone data analyzed here), i-vectors and JFA are in the same ballpark without significant differences in performance—a trend which remained until SRE10.

NIST SRE 2010 is still a very important benchmark as it is the last one (with the exception of rather special NIST SRE 2012) that evaluates purely on English data. Over time, the best published results on telephone data (condition 5) got below 1% EER. Such good results were achieved partly due to over-tuning, both of the model settings and of training data selection. This is in line with our experiments where without any special tuning and data selection, we obtained similar results as those achieved during the actual evaluation. Again, the trend of the results is as expected, but now even more pronounced—with JFA and IVEC being both equally much better than GMM-EC and XVEC bringing additional significant gain. Our single best system from SRE2010 (i-vector) has a slightly different performance than the i-vector system used here. There are, however, two important differences: in the submission, we used a full covariance UBM and we did not use the length normalization.

NIST SRE 2016 is a very challenging evaluation: participants were dealing with completely new transmission channel as the data were, for the first time, recorded outside the US. Apart from the new channel, and more importantly, the difficulty lies in two new languages of the evaluation data. Participants were exposed to only limited unlabeled data from the target domain and its correct use in score normalization and domain adaptation was crucial to obtain good results. I-vector and x-vector systems have reached a similar performance, but starting here, the x-vector systems began to dominate and the best published results achieved with more training data and more tuning towards SRE16 are well below the levels reported here. The best single system and fusion in our SRE16 submission were using unlabeled data for adaptation, while systems developed for the purposes of this comparison were neither exposed to this data nor they employed any domain adaptation, therefore their results are slightly worse. However, the effect of the adaptation on the Cantonese subset that we use here would not be as high as on the complete SRE16 test set where we would face the bi-modality of scores caused by the two languages (Cantonese and Tagalog). Reporting on Cantonese only, therefore, removes this obstacle and makes the trial set more compatible with other test sets. Characteristics of the full SRE16 are of course very interesting when exploring score normalization and domain adaptation which is not the focus of this work.

Finally, the telephone portion of NIST SRE 2018 provides an ideal picture by showing improvements from every technique. It was the first time we used x-vectors for NIST SRE and they completely outperformed i-vectors (even during system development). The better results in our submission can be explained by the utilization of domain adaptation using the provided development data.

5.2. Impact of training data size on performance

In the previous section, we analyzed the performance of various techniques as if they were known during the previous NIST SREs. There is, however, an important difference between those systems: the amount of available *labeled* training data. In Fig. 2, we analyze how individual systems change with different amounts of training data. We chose the favorite *extended condition 5* from NIST SRE 2010 and trained the systems on multiple datasets released before SRE10 as well as on Voxceleb.

It is evident that using only a large amount of out-of-domain Voxceleb data yields bad results because we introduce a channel mismatch and we do not employ any domain adaptation. The exception is partly the x-vector system which also has issues with domain mismatch, but it can already utilize the large Voxceleb dataset much better than other techniques. It should, however, be remembered that the x-vector recipe uses data augmentation. Although previous works have shown that this does not benefit i-vectors in standard evaluation set-ups (Snyder et al., 2018; Novotný et al., 2018b), we should not exclude the possibility that data augmentation could benefit i-vectors in situations with large domain mismatch. However, such an analysis is outside the scope of this paper. When training the x-vector with in-domain telephone data, it is simply not properly trained until at least SWB + SRE04 data are provided. Then it improves significantly with additional training data and obtains the best results. On the other extreme, when observing the GMM system, it is worth noticing how quickly it saturates and yields similar performance everywhere, not being able to extract additional discriminative power from more training data.

A consistent trend and a very similar performance can be observed for JFA, GMM-EC and IVEC systems. These techniques gradually improve with more data until the point of training on SWB + SRE04–06 data, where they start to saturate. Adding SRE08 into the training mix barely improves the results. In the group of these three techniques, it is interesting to notice a better resiliency of IVEC system against the domain mismatch when trained only on Voxceleb.

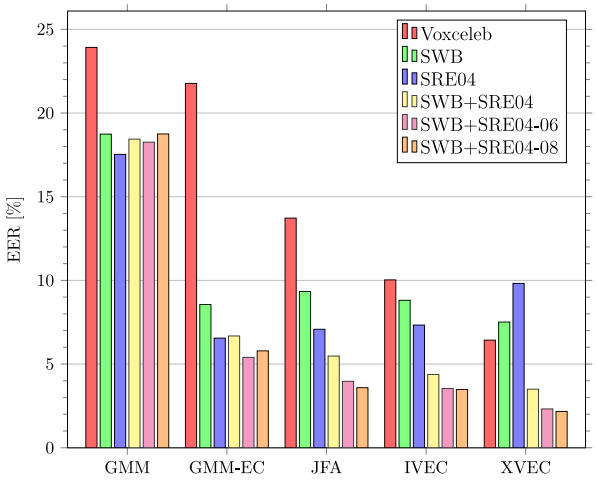

Fig. 2. Impact of amount of training data on the performance of individual techniques. All systems are evaluated on telephone condition from NIST SRE 2010 (condition 5).

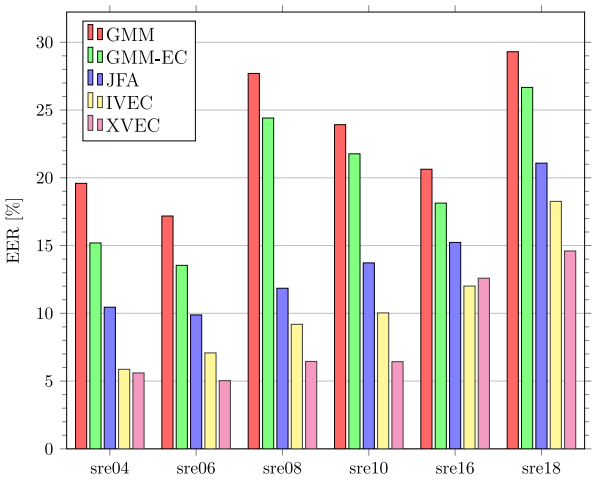

Fig. 3. Difficulty of different NIST SREs analyzed via training of all systems on completely out-of-domain Voxceleb dataset. Systems are evaluated without any domain adaptation and without exposure to any NIST data.

5.3. Increasing difficulty of NIST SREs

Having a large dataset, which is completely independent of all NIST SREs, invites for an analysis of their difficulty. We trained all systems only on the Voxceleb data and evaluated them on all NIST benchmarks. In Fig. 3, we can compare individual NIST SREs and mostly we observe a trend of increasing EER for most of the systems, which suggests also increasing difficulty. Indeed one of the biggest factors that influenced the difficulty since SRE16 is the duration of enrollment and test segments as it can be seen in Table 4. Another factor causing the increased difficulty of SRE16 and SRE18 is the nature of the test data that are non-English and collected outside of the United States. Again, we have to keep in mind that the Voxceleb dataset is completely out-of-domain, which will cause our EERs to be consistently higher, but the overall trends should remain the same.

From observations made in the previous section, we can safely ignore the results of GMM with relevance MAP. It is also difficult to make conclusions about GMM-EC as the EERs raise above 20% already for SRE08 and then for SRE10 and SRE18. The trend of increasing difficulty is however clearly visible with JFA and IVEC. Focusing our analysis to the XVEC system, we can observe a flatter trend until SRE10 suggesting that for the x-vector model, these evaluations pose a similar challenge.

At this point, it is important to mention that we provide the difficulty analysis only on the telephone data and selected conditions (or subsets) of particular NIST evaluations. We also want to point out that SRE16 would be more difficult if we considered the full test set (both Cantonese and Tagalog) as Tagalog trials on their own are harder and by evaluating both languages together, the system needs to deal with two clusters of scores which not only makes the calibration more difficult, but it also adversely affects the EER before and also after any calibration. In SRE18, this would not have been the case as the CMN2 and VAST form two distinctive conditions with different operating points and a value of the SRE18 primary metric would be an average of two separate actual costs. Considering only Cantonese trials for SRE16 and Tunisian Arabic trials for SRE18 makes our analysis easier as we avoid many factors that were specific for these two particular evaluations.

5.4. Looking forwad—wideband data

The last NIST SRE (i.e., SRE18) offered a peek into the domain of wideband audio extracted from amateur online videos by including trials derived from part of the VAST collection (Video Annotation for Speech Technologies) into the evaluation. Such a data is growing fast thanks to the massive amount of multimedia material uploaded to the internet via social networks and other channels. Hand in hand with the growth of this domain increases the interest to index this data, search in it and extract information.

Having participated in the community organized Speakers in the Wild (SITW) (Novotný et al., a; Mitchell McLaren, 2016) and Voices Obscured in Complex Environmental Settings (VOiCES) (Nandwana et al., 2019; Matějka et al., 2019) challenges, we are in a position to offer an analysis also in this domain.

Looking at Fig. 4, we can observe rather dramatic progress. The first five bars show improvements obtained from improving the SR method moving from the GMM towards x-vector while still training on now out-of-domain NIST SRE data. From this point, we stay with i-vectors and x-vectors. First, we observe the effect of even larger and now in-domain Voxceleb training dataset on i-vectors and compare working in narrow-band (by simply downsampling all of the data) with moving to wideband where the latter provides an additional performance improvement. Finally, we observe the same with x-vectors, where already the narrow-band system is significantly better than previous wideband i-vector and wideband x-vector obtaining additional improvement.

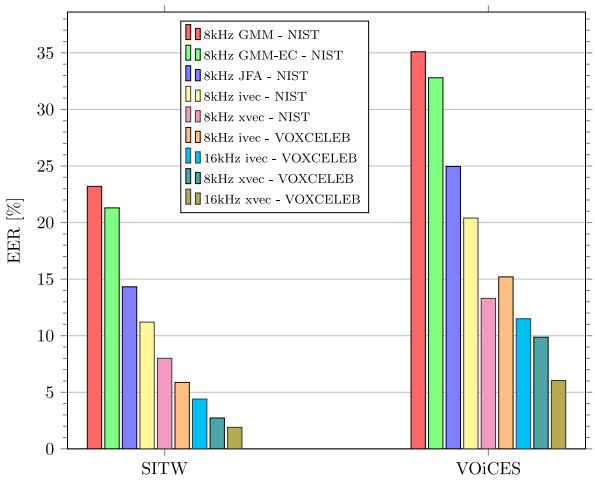

Fig. 4. Evaluation of different systems on SITW and VOiCES challenges.

The trend of decreasing EER is the same for both SITW and VOiCES datasets where the latter is more challenging being designed to reflect mostly noisy and reverberant data.

6. Conclusion

In this paper, we have provided a comprehensive analysis and historical overview of main speaker recognition methods all of which were considered state-of-the-art at certain periods of time; in the case of x-vectors, we worked with the current state-of-the-art SR technique. The presented analysis of all techniques on multiple NIST SREs, SITW and VOiCES allows us not only to directly compare historical methods but also to observe many interesting trends that were revealed during a large experimental effort.

Multiple times, we have seen a difference in performance of GMM with relevance MAP, GMM with eigen-channel compensation, JFA, i-vectors, and x-vectors with PLDA on various benchmarks. We have analyzed how the amount of speaker-labeled training data impacts the performance of these techniques and then we have analyzed the NIST SRE's themselves in terms of difficulty where we found that indeed the difficulty has been generally an increasing trend. We have also glimpsed at non-NIST evaluations to see the trends in a new domain of wideband data that are increasingly available as uploads to social networks or other online sources.

Acknowledgments

The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", Google Faculty Research Award program, Czech National Science Foundation (GACR) project No. GJ17-23870Y, Czech National Science Foundation (GACR) project "NEUREM3" No. 19-26934X, Marie Sklodowska-Curie grant agreement No. 748097, the Marie Sklodowska-Curie cofinanced by the South Moravian Region under grant agreement No. 665860, Technology Agency of the Czech Republic project No. TJ01000208 "NOSICI", and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science"—LQ1602.

References

Alam, J., Bhattacharya, G., Brummer, N., Burget, L., Diez, M., Glembek, O., Kenny, P., Klčo, M., Landini, N.F., Lozano, A.D., Matějka, P., Monteiro, J., Mošner, L., Novotný, O., Plchot, O., Profant, J., Rohdin, A.J., Silnova, A., Slavíček, J., Stafylakis, T., Zeinali, H., 2018. Abc NIST SRE 2018 system description. In: Proceedings of NIST SRE Workshop. National Institute of Standards and Technology.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-Independent speaker verification systems. Digit. Signal Process. 10 (1), 42–54.

Beneš, K., Kesiraju, S., Burget, L., 2018. I-vectors in language modeling: An efficient way of domain adaptation for feed-forward models. In: Proceedings of Interspeech 2018, pp. 3383–3387.

Brummer, N., 2004. Spescom DataVoice NIST 2004 system description. In: Proceedings of the NIST Speaker Recognition Evaluation.

Brummer, N., Burget, L., Kenny, P., Matejka, P., de Villiers, E., Karafiat, M., Kockmann, M., Glembek, O., Plchot, O., Baum, D., Senoussauoi, M., 2010. ABC System description for NIST SRE. In: Proceedings of the NIST Speaker Recognition Evaluation. Brno University of Technology, pp. 1–20.

Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van, D.L., Matějka, P., Schwarz, P., Strasheim, A., 2007. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation. IEEE Trans Audio Speech Lang Process.

Burget, L., Fapšo, M., Hubeika, V., Glembek, O., Karafiát, M., Kockmann, M., Matějka, P., Schwarz, P., Černocký, J., 2009. BUT system for NIST speaker recognition evaluation. In: Proceedings of the Interspeech.

Burget, L., Matějka, P., Hubeika, V., Černocký, J., 2009. Investigation into variants of Joint Factor Analysis for speaker recognition. In: Proceedings of the Interspeech.

Burget, L., Matejka, P., Schwarz, P., Glembek, O., Cernocky, J., 2007. Analysis of feature extraction and channel compensation in GMM speaker recognition system. IEEE Trans. Audio Speech Lang. Process. 15 (7), 1979–1986.
Burget, L., Plchot, O., Cumani, S., Glembek, O., Matějka, P., Brümmer, N., 2011. Discriminatively trained Probabilistic Linear Discriminant Analysis for Speaker Verification. In: Proceedings of the ICASSP. Prague, CZ.
Campbell, J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85 (9), 1437–1462.
Cumani, S., Brümmer, N., Burget, L., Laface, P., 2011. Fast Discriminative Speaker Verification in the I-Vector Space. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, pp. 4852–4855.
Cumani, S., Brummer, N., Burget, L., Laface, P., Plchot, O., Vasilakakis, V., 2013. Pairwise discriminative speaker verification in the I -Vector space. IEEE Trans. Audio Speech Lang Process. 2013 (6), 1217–1227.
Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., Dumouchel, P., 2009. Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In: Proceedings of the Interspeech. Brighton, UK.
Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2010. Front-End factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. PP (99), 1. https://doi.org/10.1109/TASL.2010.2064307.
Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–38.
Diez, M., Burget, L., Matějka, P., 2018. Speaker Diarization based on Bayesian HMM with Eigenvoice Priors. In: Proceedings of the Odyssey the Speaker and Language Recognition Workshop.
Fér, R., Matějka, P., Grézl, F., Plchot, O., Černocký, J., 2015. Multilingual bottleneck features for language recognition. Proceedings of the Interspeech.
Ferrer, L., Bratt, H., Burget, L., Cernocky, H., Glembek, O., Graciarena, M., Lawson, A., Lei, Y., Matejka, P., Plchot, O., et al., 2011. PRISM Evaluation Set Promoting Robustness for Speaker Modeling in the Community.
Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of I-vector Length Normalization in Speaker Recognition Systems. In: Proceedings of the Interspeech. Florence, Italy.
Garcia-Romero, D., McCree, A., 2014. Supervised domain adaptation for i-vector based speaker recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4047–4051.
Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2, 291–298.
Glembek, O., 2012. Optimization of Gaussian Mixture Subspace Models and Related Scoring Algorithms in Speaker Verification. Brno University of Technology, Faculty of Information Technology.
Glembek, O., Burget, L., Brümmer, N., Plchot, O., Matějka, P., 2011. Discriminatively Trained i-vector Extractor for Speaker Verification. In: Proceedings of the Interspeech, pp. 137–140.
Glembek, O., Burget, L., Dehak, N., Brummer, N., Kenny, P., 2009. Comparison of scoring methods used in speaker recognition with Joint Factor Analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4057–4060.
Glembek, O., Burget, L., Matejka, P., Karafiat, M., Kenny, P., 2011. Simplification and optimization of I-Vector Extraction. In: Proceedings of the ICASSP.
Glembek, O., Ma, J., Matějka, P., Zhang, B., Plchot, O., Burget, L., Matsoukas, S., 2014. Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems. In: Proceedings of the ICASSP.
Graff, D., Canavan, A., Zipperlen, G., 1998. Switchboard-2 phase I. Linguistic Data Consortium, Philadelphia.
Graff, D., Miller, D., Walker, K., 2001. Switchboard cellular part 1 audio. Linguistic Data Consortium, Philadelphia.
Graff, D., Miller, D., Walker, K., 2002. Switchboard-2 phase III. Linguistic Data Consortium, Philadelphia.
Graff, D., Miller, D., Walker, K., 2004. Switchboard cellular part 2 audio. Linguistic Data Consortium, Philadelphia.
Graff, D., Walker, K., Canavan, A., 1999. Switchboard-2 phase II. Linguistic Data Consortium, Philadelphia.
Greenberg, C.S., Stanford, V.M., Martin, A.F., Yadagiri, M., Doddington, G.R., Godfrey, J.J., Hernandez-Cordero, J., 2013. The NIST 2012 speaker recognition evaluation. In: Proceedings of the International Conference on Spoken Language Processing (Interspeech).
Grézl, F., Karafiát, M., Kontár, S., Černocký, J., 2007. Probabilistic and bottle-neck features for LVCSR of meetings. In: Proceedings of the ICASSP.
Hermansky, H., Morgan, N., 1994. RASTA Processing of speech. IEEE Trans. Speech Audio Process.
Ioffe, S., 2006. Probabilistic Linear Discriminant Analysis. In: Proceedings of the ECCV (4), pp. 531–542.
Kajarekar, S.S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L., Bocklet, T., 2009. THE SRI NIST speaker recognition evaluation system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4205–4208.
Kenny, P., 2010. Bayesian speaker verification with heavy-tailed priors. In: Proceedings of the Odyssey the Speaker and Language Recognition Workshop.
Kenny, P., Dumouchel, P., 2004. Disentangling speaker and channel effects in speaker verification. In: Proceedings of the ICASSP.
Kenny, P., Mihoubi, M., Dumouchel, P., 2003. New MAP Estimators for Speaker Recognition. In: Proceedings of the Eurospeech.
Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of inter-Speaker variability in speaker verification. IEEE Trans. Audio, Speech Lang. Process. 16 (5), 980–988.
Kesiraju, S., Burget, L., Szőke, I., Černocký, J., 2016. Learning document representations using subspace multinomial model. In: Proceedings of Interspeech, pp. 700–704.
Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. 52 (1), 12–40.
Kockmann, M., 2012. Subspace Modeling of Prosodic Features for Speaker Verification. Brno University of Technology, Faculty of Information Technology.
Kumar, N., 1997. Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition. John Hopkins University, Baltimore.
Lei, Y., Scheffer, N., Ferrer, L., McLaren, M., 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In: Proceedings of the ICASSP.
Lozano, A.D., Silnova, A., Matějka, P., Glembek, O., Plchot, O., Pešán, J., Burget, L., Gonzalez-Rodriguez, J., 2016. Analysis and Optimization of Bottleneck Features for Speaker Recognition. In: Proceedings of Odyssey.
Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The det curve in assessment of detection task performance. pp. 1895–1898.
Martin, A., Greenberg, C., 2009. NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels. In: Proceedings of the International Conference on Spoken Language Processing (Interspeech).
Martin, A., Greenberg, C., 2010. The NIST 2010 speaker recognition evaluation. In: Proceedings of the International Conference on Spoken Language Processing (Interspeech).
Matějka, P., Burget, L., Schwarz, P., Glembek, O., Karafiát, M., Grézl, F., Černocký, J., van, D. L., Brümmer, N., Strasheim, A.,. STBU system for the NIST speaker recognition evaluation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Matějka, P., Glembek, O., Castaldo, F., Alam, J., Plchot, O., Kenny, P., Burget, L., Černocký, J., 2011. Full-covariance UBM and Heavy-tailed PLDA in I-Vector Speaker Verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP.
Matějka, P., Glembek, O., Novotný, O., Plchot, O., Grézl, F., Burget, L., Černocký, J., 2016. Analysis Of DNN Approaches To Speaker Identification. In: Proceedings of ICASSP.
Matějka, P., Novotný, O., Plchot, O., Burget, L., Diez, M.S., Černocký, J., 2017. Analysis of Score Normalization in Multilingual Speaker Recognition. In: Proceedings of Interspeech. International Speech Communication Association, pp. 1567–1571.
Matějka, P., Plchot, O., Zeinali, H., Mošner, L., Silnova, A., Burget, L., Novotný, O., Glembek, O., 2019. Analysis of BUT submission in far-field scenarios of VOiCES 2019 challenge. In: Proceedings of the Interspeech.
Mitchell McLaren, D.C.A.L., Ferrer, L., 2016. The speakers in the wild (sitw) speaker recognition database. In: Proceedings of the Interspeech.

Mošner, L., Plchot, O., Matějka, P., Novotný, O., Černocký, J., 2018. Dereverberation and Beamforming in Robust Far-Field Speaker Recognition. In: Proceedings of Interspeech. International Speech Communication Association, pp. 1334–1338.
National Institute of Standard and Technology, http://www.nist.gov/speech/tests/spk/index.htm.
Nagrani, A., Chung, J.S., Zisserman, A., 2017. Voxceleb: A large-scale speaker identification dataset. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association Interspeech, Stockholm, Sweden, August 20−24, 2017, pp. 2616–2620.
Nandwana, M. K., van Hout, J., McLaren, M., Lawson, A., Barrios, M. A., 2019. The VOiCES from a distance challenge 2019 evaluation plan. In: arXiv:1902.10828 [eess.AS].
Ng, T., Zhang, B., Nguyen, L., Matsoukas, S., Zhou, X., Mesgarani, N., Veselý, K., Matějka, P., 2012. Developing a Speech Activity Detection System for the DARPA RATS Program. In: Proceedings of Interspeech, pp. 1–4.
Novotný, O., Matějka, P., Glembek, O., Plchot, O., Grézl, F., Burget, L., Černocký, J., 2016. Analysis of the DNN-Based SRE Systems in Multi-language Conditions. In: Proceedings of the SLT, pp. 199–204.
Novotný, O., Matějka, P., Plchot, O., Glembek, O., Burget, L., Černocký, J., a. Analysis of Speaker Recognition Systems in Realistic Scenarios of the SITW 2016 Challenge. In: Proceedings of Interspeech 2016, pp. 828−832.
Novotný, O., Plchot, O., Matějka, P., Mošner, L., Glembek, O., 2018b. On the use of X-vectors for Robust Speaker Recognition. In: Proceedings of the Odyssey.
Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association INTERSPEECH, Dresden, Germany, September 6−10, 2015, pp. 3214–3218.
Plchot, O., 2014. Extensions to Probabilistic Linear Discriminant Analysis for Speaker Recognition. Brno University of Technology, Faculty of Information Technology.
Plchot, O., Matějka, P., Silnova, A., Novotný, O., Diez, M., Rohdin, A.J., Glembek, O., Brümmer, N., Swart, A., Prieto, J.J., Garcia, P.L.P., Buera, L., Kenny, P., Alam, J., Bhattacharya, G., 2019. Analysis and description of ABC submission to NIST SRE 2016. In: Proceedings of the Interspeech. International Speech Communication Association.
Plchot, O., Matsoukas, S., Matějka, P., Dehak, N., Ma, J., Cumani, S., Glembek, O., Heřmanský, H., Mesgarani, N., Soufifar, M.M., Thomas, S., Zhang, B., Zhou, X., et al., 2013. Developing A Speaker Identification System For The DARPA RATS Project. In: Proceedings of ICASSP, pp. 6768–6772.
Prince, S.J.D., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proceeding of the International Conference on Computer Vision (ICCV). Rio de Janeiro, Brazil.
Reynolds, D.A., 2003. Channel robust speaker verification via feature mapping. In: Proceedings of the ICASSP.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Digit Signal Process.
Rohdin, J., Silnova, A., Diez, M.S., Plchot, O., Matějka, P., Burget, L., 2018. End-to-End DNN Based Speaker Recognition Inspired by i-Vector and PLDA. In: Proceedings of the ICASSP, pp. 4874–4878.
Rohdin, J., Stafylakis, T., Silnova, A., Zeinali, H., Burget, L., Plchot, O., 2019. Speaker verification using end-to-end adversarial language adaptation. In: Proceedings of the ICASSP, pp. 6006–6010.
Sadjadi, S.O., Greenberg, C.S., Singer, E., Reynolds, D.A., Mason, L.P., Hernandez-Cordero, J., 2019. The NIST 2018 speaker recognition evaluation. In: Proceedings of the International Conference on Spoken Language Processing (Interspeech).
Sadjadi, S.O., Kheyrkhah, T., Tong, A., Greenberg, C.S., Reynolds, D.A., Singer, E., Mason, L.P., Hernandez-Cordero, J., 2017. The NIST 2016 speaker recognition evaluation. In: Proceedings of the International Conference on Spoken Language Processing (Interspeech).
Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S., 2017. Deep neural network Embeddings for text-independent speaker verification. Proc. Interspeech 999–1003.
Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., Khudanpur, S., 2019. Speaker recognition for multi-speaker conversations using x-vectors. In: Proceedings of the ICASSP.
Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-Vectors: robust DNN embeddings for speaker recognition. Proc. ICASSP.
Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., Khudanpur, S., 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 165–170.
Solomonoff, A., Quillen, C., Campbell, W.M., 2004. Channel compensation for SVM speaker recognition. In: Proceedings of the Odyssey-04, Speaker and Language Recognition Workshop, pp. 57−62.
Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A., 2005. MLLR transforms as features in speaker recognition. In: Proceedings of the 9th European Conference on Speech Communication and Technology, pp. 2425–2428.
Sturim, D., Campbell, W., N. Karam, Z., Reynolds, D., Richardson, F., 2009. The MIT lincoln laboratory 2008 speaker recognition system.pp. 2359−2362.
Veselý, K., Karafiát, M., Grézl, F., Janda, M., Egorova, E., 2012. The Language-Independent Bottleneck Features. In: Proceedings of IEEE Workshop on Spoken Language Technology, pp. 336−341.
Villalba, J., Lleida, E., 2014. Unsupervised adaptation of PLDA by using variational Bayes methods. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 744–748.
Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J., 1989. Phoneme recognition using time-delay neural networks. In: Proceedings of the IEEE Transactions on Acoustics, Speech, and Signal Processing, 37, pp. 328–339.