We are happy to announce the release of the MSR Identity Toolbox: A MATLAB toolbox for speaker-recognition research. This toolbox contains a collection of MATLAB tools and routines that can be used for research and development in speaker recognition. It provides researchers with a test bed for developing new front-end and back-end techniques, allowing replicable evaluation of new advancements. It will also help newcomers in the field by lowering the "barrier to entry," enabling them to quickly build baseline systems for their experiments. Although the focus of this toolbox is on speaker recognition, it can also be used for other speech related applications such as language, dialect, and accent identification. Additionally, it provides many of the functionalities available in other open-source speaker recognition toolkits (e.g., ALIZE [1]) but with a simpler design which makes it easier for the users to understand and modify the algorithms.

The MATLAB tools in the Identity Toolbox are computationally efficient for three reasons: vectorization, parallel loops, and distributed processing. First, the code is simple and easy for MATLAB to vectorize. With long vectors, most of the CPU time is spent in optimized loops, which are the core of the processing. Second, the code is designed for parallelization available through the Parallel Computing Toolbox (i.e., the toolbox codes use "parfor" loops). Without the Parallel Computing Toolbox, these loops execute as normal "for" loops on a single CPU. But when this toolbox is installed, the loops are automatically distributed across all the available CPUs. In our pilot experiments, the codes were run across all 12 cores in a single machine. Finally, the primary computational routines are designed to work as compiled programs. This makes it easy to distribute the computational work to all the machines on a computer cluster, without the need for additional licenses.

SPEAKER ID BACKGROUND

In recent years, the design of robust and effective speaker-recognition algorithms has attracted significant research effort from academic and commercial institutions. Speaker recognition has evolved substantially over the past few decades; from discrete vector quantization (VQ) based systems [2] to adapted Gaussian mixture model (GMM) solutions [3], and more recently to factor analysis based Eigenvoice (i-vector) frameworks [4]. The Identity Toolbox, version 1.0, provides tools that implement both the conventional GMM-UBM and state-of-the-art i-vector based speaker-recognition strategies.

As shown in Fig. 1, a speaker-recognition system includes two primary components: a front-end and a back-end. The
front-end transforms acoustic waveforms into more compact and less redundant acoustic feature representations. Cepstral features are most often used for speaker recognition. It is practical to only retain the high signal-to-noise ratio (SNR) regions of the waveform, therefore there is also a need for a speech activity detector (SAD) in the front-end. After dropping the low SNR frames, acoustic features are further post-processed to remove the linear channel effects. Cepstral mean and variance normalization (CMVN) \[5\] is commonly used for the post-processing. The CMVN can be applied globally over the entire recording or locally over a sliding window. Feature warping \[6\], which is also applied over a sliding window, is another popular feature-normalization technique that has been successfully applied for speaker recognition. This toolbox provides support for these normalization techniques, although no tool for feature extraction or SAD is provided. The Auditory Toolbox \[7\] and VOICEBOX \[8\], which are both written in MATLAB, can be used for feature extraction and SAD purposes.

The main component of every speaker-recognition system is the back-end where speakers are modeled (enrolled) and verification trials are scored. The enrollment phase includes estimating a model that represents (summarizes) the acoustic (and often phonetic) space of each speaker. This is usually accomplished with the help of a statistical background model from which the speaker-specific models are adapted. In the conventional GMM-UBM framework the universal background model (UBM) is a Gaussian mixture model (GMM) that is trained on a pool of data (known as the background or development data) from a large number of speakers \[3\]. The speaker-specific models are then adapted from the UBM using the maximum a posteriori (MAP) estimation. During the evaluation phase, each test segment is scored either against all enrolled speaker models to determine who is speaking (speaker identification), or against the background model and a given speaker model to accept/reject an identity claim (speaker verification).

On the other hand, in the i-vector framework the speaker models are estimated through a procedure called Eigenvoice adaptation \[4\]. A total variability subspace is learned from the development set and is used to estimate a low (and fixed) dimensional latent factor called the identity vector (i-vector) from adapted mean supervectors (the term "i-vector" sometimes also refers to a vector of "intermediate" size, bigger than the underlying cepstral feature vector but much smaller than the GMM supervector). Unlike the GMM-UBM framework, which uses acoustic feature vectors to represent the test segments, in the i-vector paradigm both the model and test segments are represented as i-vectors. The dimensionality of the i-vectors are normally reduced through linear discriminant analysis (with Fisher criterion \[9\]) to annihilate the non-speaker related directions (e.g., the channel subspace), thereby increasing the discrimination between speaker subspaces. Before modelling the dimensionality-reduced i-vectors via a generative factor analysis approach called the probabilistic LDA (PLDA) \[10\], they are mean and length normalized. In addition, a whitening transformation that is learned from i-vectors in the development set is applied. Finally, a fast and linear strategy \[11\], which computes the log-likelihood ratio (LLR) between same versus different speaker's hypotheses, scores the verification trials.

**IDENTITY TOOLBOX**

The Identity toolbox provides tools for speaker recognition using both the GMM-UBM and i-vector paradigms. It has been attempted to maintain consistency with the naming convention in the code to follow the formulation and symbolization used in the literature. This will make it easier for the users to compare the theory with the implementation and help them better understand the concept behind each algorithm. The tools can be run from a MATLAB command line using available parallelization (i.e., parfor loops), or compiled and run on a computer cluster without the need for a MATLAB license. The toolbox includes two demos which use artificially generated features to show how different tools can be combined to build and run GMM-UBM and i-vector based speaker recognition systems. In addition, the toolbox contains scripts for performing a small-scale speaker identification experiment using the TIMIT database. Moreover, we have replicated state-of-the-art results on the large-scale NIST SRE-2008 core tasks (i.e., short2-short3 conditions \[15\]). The list below shows the different tools available in the toolbox, along with a short descriptions of their capabilities:

**Feature normalization**

- Global cepstral mean and variance normalization (cmvn)
IEEE SLTC Newsletter, November 2013

- Cepstral mean and variance normalization over a sliding window (wcmvn)
- Short-term Gaussianization (a.k.a. feature warping) over a sliding window (fea_warping) [6]

**GMM-UBM**
- Gaussian mixture model (GMM) learning using expectation maximization (gmm-em) [3]
- GMM adaptation using maximum a posteriori estimation (mapAdapt) [3]
- GMM-based verification trial scoring (score_gmm_trials) [3]

**i-vector-PLDA**
- Sufficient statistics computation for observations given the GMM (compute_bw_stats) [4, 12]
- Total variability subspace learning using EM (train_tv_space) [4, 12, 13]
- i-vector extraction (extract_ivector) [4, 12, 13]
- Linear discriminant analysis (lda) [9]
- i-vector length normalization, centering, whitening, and Gaussian probabilistic LDA using EM (gplda-em) [10, 11, 14]
- PLDA-based verification trial scoring (score_gplda_trials) [11, 14]

**EER and DET plot**
- Equal error rate (EER), detection cost function (DCF), and detection error tradeoff (DET) (compute_eer) [15, 16]

The Identity Toolbox is available from the MSR website (http://research.microsoft.com/jump/203749) under a Microsoft Research License Agreement (MSR-LA) that allows use and modification of the source codes for non-commercial purposes. The MSR-LA, however, does not permit distribution of the software or derivative works in any form.

**REFERENCES**


http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html


Seyed Omid Sadjadi is a PhD candidate at the Center for Robust Speech Systems (CRSS), The University of Texas at Dallas. His research interests are speech processing and speaker identification. Email: omid.sadjadi@ieee.org

Malcolm Slaney is a Researcher at Microsoft Research in Mountain View, California, a Consulting Professor at Stanford CCRMA, and an Affiliate Professor in EE at the University of Washington. He doesn’t know what he wants to do when he grows up. Email: malcolm@ieee.org

Larry Heck is a Researcher in Microsoft Research in Mountain View California. His research interests include multimodal conversational interaction and situated NLP in open, web-scale domains. Email: larry.heck@ieee.org