Project Title: **TASRE** *(TigerIT Automatic Speaker Recognition Engine)*

2021 NIST System Description Report: Report

Overview

● Support for 1 to 1 matching and 1 to N verification.
● Proprietary silence detection, noise removal, and speech feature extraction techniques were developed in our R&D lab.
● Robust speech features a templating and matching system.
● An optimised parallel matching algorithm that scales up for both CPU and GPU resources.
● Minimal setup cost.
● Easy to integrate with an existing biometric identification system.
● Live and offline enrollment from diverse sources.
● Batch enrollment facility.

System Architecture

● The system collects speech recording and related personal information from enrollment sessions and stores all information in the system database. The database can be hosted both in the cloud and the client-server.
● The state-of-the-art matching algorithm is equipped with fast and parallel searching and scalable in both CPU and GPU.
● Matching requests can come from registration centres (to identify if an applicant was enrolled previously) or from dedicated identity-checking points (to verify the identity of a registered person).
● A matching request is sent from the remote station along with the collected speech recording. Our speech matching solution receives the request, performs matching, and replies with the result to the respective station.

● The central server also has adjudication and administration stations which are used to perform matching operations centrally and manage the overall server as needed. The solution also uses disaster recovery servers to keep a backup copy of database content.

Process

● Voice data is segmented using advanced silence detection and noise reduction tools.
● Channel data is processed according to our requirements.
● Data is resampled to as per our algorithm requirement.
● Extraction of frontend features for speaker recognition.
● Proprietary features along with our modified x-vector features are extracted from the available pre-processed audio sessions. Finally, our proprietary DNN embeddings are stored for matching and searching purposes.

Technical specification

● Our system expects speech of 16KHz sampling rate, at least 16bit depth, single-channel audio.
● Our system can enrol speech with only 5 seconds of speech data (without any noise or silence) in a single session at a bare minimum. However, an increased amount of audio with multiple sessions will improve the recognition quality.
● We use a tensor-based file format for efficient storing and management of embedding vectors. Each template takes around 6KB of disk space. However, the initial disk space required by each audio sample is varied by speech length.
● Template extraction time using DNN depends on speech length.

Experimental Results

Our speaker recognition system is completely text-independent. Here, the recognition process is not dependent on what the user is speaking. We have performed multiple separate experiments on different standard

datasets to evaluate the performance of our algorithm. Of them, three are presented below.

Experiment 1 used voxceleb-1 test database. Experiment 2 used the TIMIT speaker dataset and finally, in the third experiment, our private multi-lingual south-Asian speaker dataset was used. All of the speech samples are unique and unseen to the model.

	Experiment 1	Experiment 2	Experiment 3
Dataset	Voxceleb-1 test (official)	TIMIT (full test dataset)	SOTW (proprietary, multi-lingual [Pashto/Dari, Bangla, Hindi, Urdu, Arabic])
Total voice samples in the dataset	4874	2100	1168
Total subjects in the database	40	168	48
Number of trials	37720	550000	1379859
EER (no norm., no calib.)	1.20%	0.40%	2.18%
minDCF (no norm., no calibration)	0.067%	0.029%	0.084%

Note that, for all of the experiments a single model is used for scoring without any post-processing (normalisation or calibration). For the datasets, the score distribution and ROC plots are demonstrated in the following section.

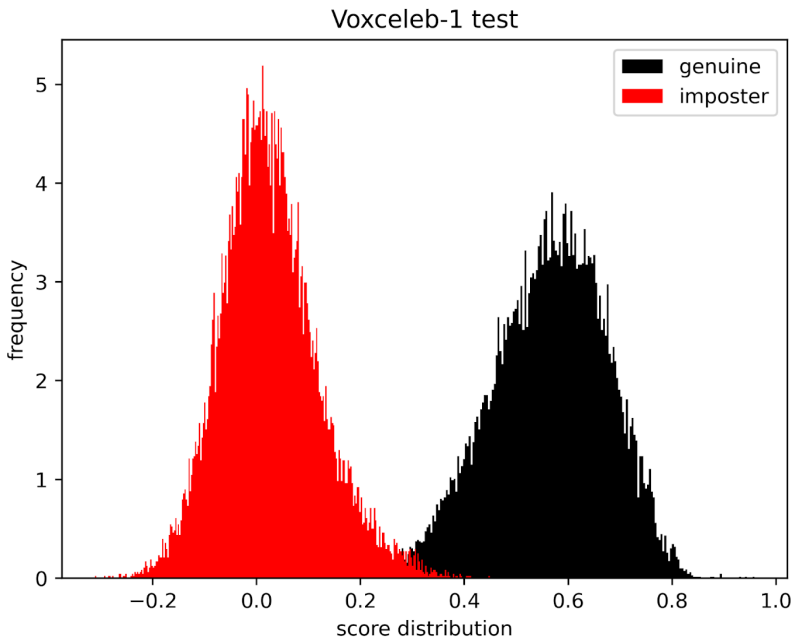

Figure 1: Model score distribution for voxceleb-1 test set

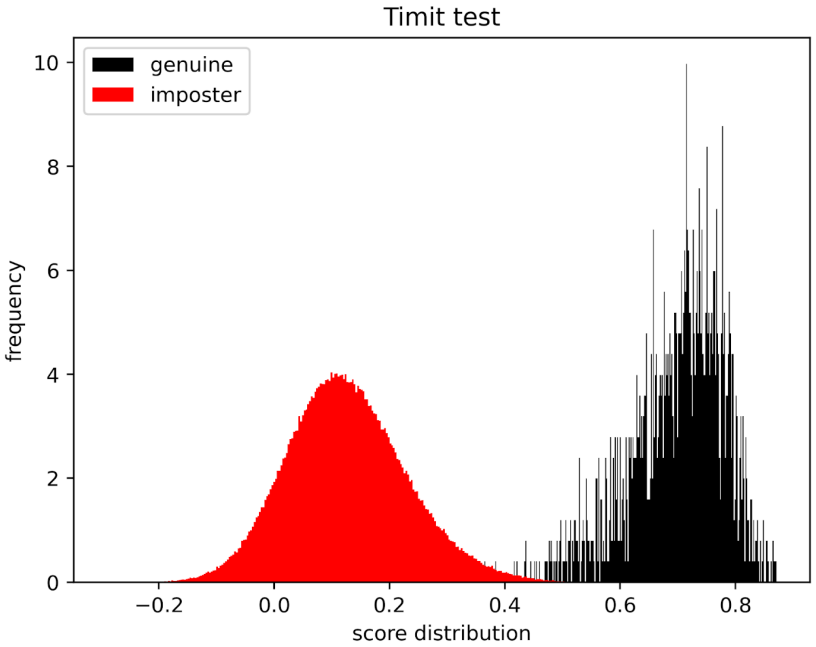

Figure 2: Model score distribution for TIMIT test set

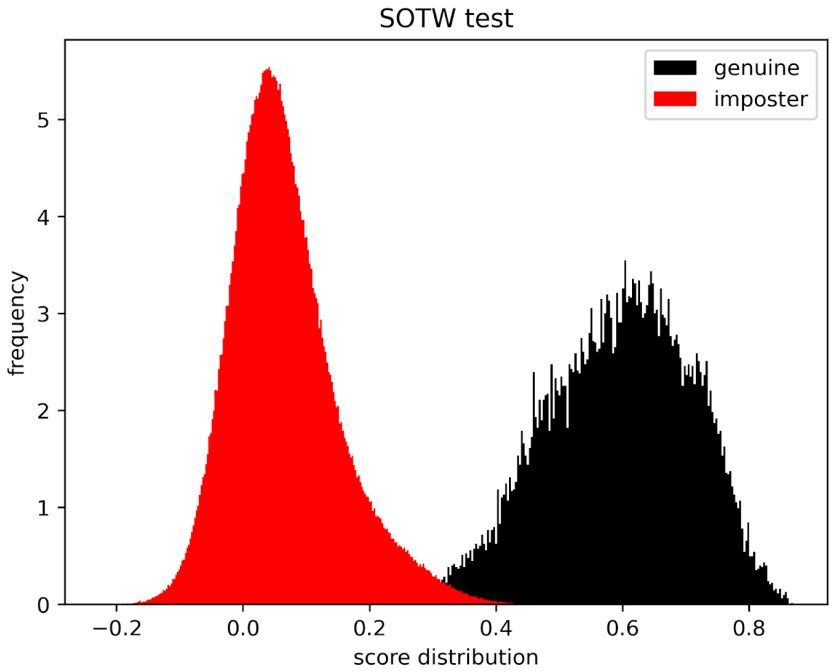

Figure 3: Model score distribution for SOTW test set

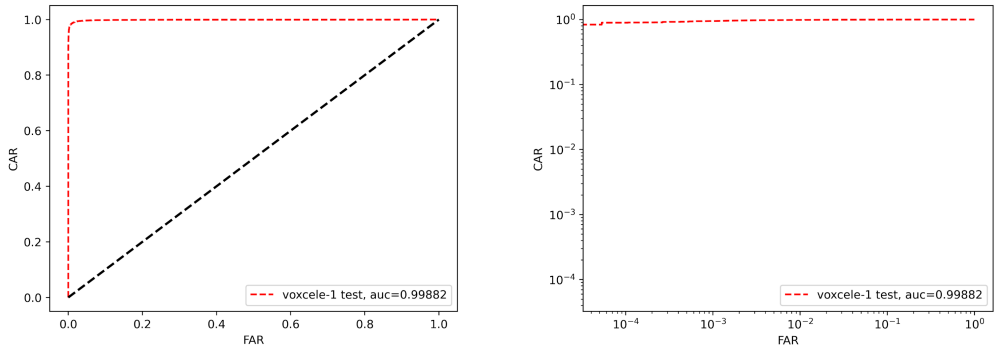

Figure 4: ROC curve for voxceleb-1 test

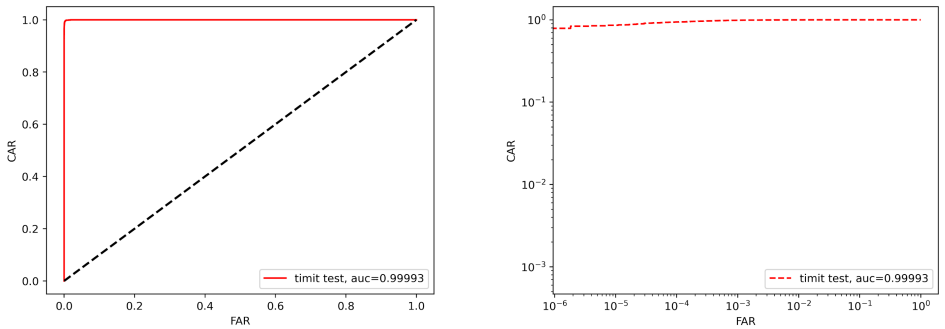

Figure 5: ROC curve for TIMIT test

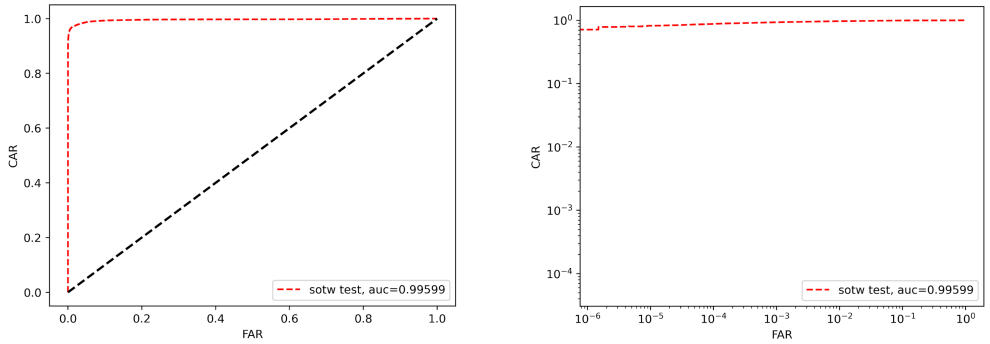

Figure 6: ROC curve for SOTW test