GIAPSI-UPM System Description for the Text-Dependent SdSV Challenge

A. Álvarez-Marquina

ETSI Informáticos, Universidad Politécnica de Madrid, SPAIN

aalvarez@fi.upm.es

Abstract

The present paper describes the main design decisions regarding the system components and several of the main aspects of its performance. The paper is structured following the guidelines contained in the Short-duration Speaker Verification (SdSV) Challenge 2020.

1. Introduction

The following text contains four different sections. Section 1 presents a complete description of the system components, including front-end and back-end modules, along with their configurations. Section 2 includes a description of the data partitions used to train the various models. Sections 3 presents the performance of the primary submitted system on the SdSV 2020 Task 1 Development set. Finally, section 4 presents a report of the CPU (single threaded) execution times as well as the amount of memory used to process a single trial.

2. Description of System Components

2.1. Front-end

2.1.1. VAD

A simple energy VAD is applied to every single input sample. A frame with less than 30dB below the maximum energy level will be considered as speech. Besides, a minimum threshold of 50dB is considered.

2.1.2. Features

The set of features is divided into 2 main groups. The first group comprises 20+20Δ+20ΔΔ MFCC features calculated for a window of 512 samples at a sampling frequency of 8kHz. A sliding window of 128 samples is always applied. A second group is comprised of the several voice quality measures, see [1], calculated for different time intervals. Time intervals are ranged from 25ms to 5000ms. These measures are estimated frame by frame but the mean and standard deviation values, corresponding to the time interval of interest, are finally extracted. For a given frame, 60+840 features are extracted.

2.2. Back-end

2.2.1. Background model

A 128 diagonal Gaussian UBM is trained using only the 60 MFCC features. The UBM will be used as a soft decision vector quantizer.

2.2.2. Feature selection procedure

In order to select useful voice quality features sets, input vectors will be associated to different Gaussians. The naive clusters are going to be built with vector that are assigned a probability higher than 0.2, so that a single vector may be associated to more than a single cluster (Gaussian).
The feature selection procedure is based on the J-measure applied to the scatter matrix [2]. The goal is to maximize the inter-speaker variance at the same time the intra-speaker variance is reduced. For every single cluster, a given number of features from the whole set is selected as the real measures for the rest of the processing. The total amount of selected features is fixed to 100.

2.2.3. Replacement UBM

A new 128 diagonal Gaussian UBM is calculated using the output probabilities from the original UBM but estimating new mean and diagonal covariance matrices by replacing MFCC by a new subset set of selected features.

2.2.4. Pseudo i-vector extractor

Pseudo i-vector components are estimated directly, that is, no T-matrix is estimated. The pseudo i-vector is the resulting of first order Baum-Welch extractor when applied to the second (replacement UBM). As a result, input feature sequences corresponding to a given utterance will be transformed into a single 128x100 element-supervector.

2.2.5. LDA

A standard LDA procedure is applied in order to reduce the dimensionality from 12800 to 512 components.

2.2.6. PLDA

A standard PLDA procedure is applied to the previous LDA vectors, maintaining the vector dimensionality.

3. Data Partitions used to Train the Models

Training materials comprise only SdSV 2020 Task-1 training speech utterances.

All the training material is divided into 3 different subsets. The first one in applied for obtaining the UBM. The second is devoted to the estimation of the voice quality feature selection stage. Finally, the third one is used in the LDA and PLDA transformations.

4. Performance of the Submitted System on the SdSV 2020 Evaluation Set

Table 1 summarizes the performance achieved. The results are taken from the Leaderboard figures.

Table 1: *Performance achieved by the primary system.*

EER (%)	MIN_DCF
13.49	0.4830

5. CPU, GPU and Memory Report

5.1. CPU report

The required CPU (single threaded) computational power for the different steps may be found in Table 2. System CPU associated to the figures corresponds to an Intel i9-9900X. Figures for building a speaker model and testing a trial are quite similar since Feature extraction stage is by a large amount the more consuming procedure. Additionally, model building needs an averaging of several LDA vectors but at the same time does not involve the execution of PLDA and Scoring modules.

Table 2: Intel *i9-9900X (single threaded) execution times.*

Module name	**RT (%)**
VAD	0.01
Feature extractor	76.50
UBM probabilities	0.10
First order Baum-Welch statistics (Replacement UBM)	2.15
LDA	0.1
PLDA	0.3
Scoring	0.01
TOTAL	***79,17***

5.2. GPU report

No GPU is required for the model estimation procedure.

5.3. Memory requirements

The global amount of memory required for the processing of a trial or building a model is ruled by the feature extraction scheme. The actual amount is 180MB. However, the implementation of that module has not been optimized form the point of view of memory consumption.

6. References

[1] Soo J. P., Yeung G., Kreiman J., Keating, P. A., and Alwan, A., “Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems”, *INTERSPEECH 2017*, August 20–24, 2017, Stockholm, Sweden

[2] Theodoridis, S. and Koutroumbas, K., *Pattern Recognition*, 4th edition, Elsevier 2009.