# Team61 SDSV Challenge 2020 Task1 System Description

*Team61*

Anonymous

anonymous

## Abstract

This is the system description for text dependent speaker verification in SDSV Challenge 2020 Task1. We propose a system consisting of sentence-specific JFA models and gender detectors combined with a speech-recognition module specifically trained for this task.

**Index Terms**: text-dependent speaker verification, joint factor analysis, speech recognition

## 1. Introduction

In text-dependent speaker verification Joint Factor Analysis [1] is a commonly used and robust approach because of its effectiveness in a limited feature space with similar features for each utterance. In this task our core system is sentence-specific JFA models and gender detectors trained for 10 sentences in the dataset separately. Scores are calculated for corresponding sentence model with applied ZT-norm [2] using corresponding gender-dependent cohort sets for each trial. For target-wrong trials, a modified speech recognition module is trained and target-wrong trial scores are detected and decreased by this module.

## 2. Dataset

DeepMine Dataset [3] Task 1 [4] Train partition is used in tranining of JFA models. This dataset consists of 963 speakers and 10 different phrase utterances by those speakers. Dataset is split to tranining and development set as utterances from first 900 speakers constitutes tranining set, utterances from last 63 speakers constitute development set.

In order to have a consistent test conditions with the evaluation set, development trials are generated using development set according to the enrollment and authentication scheme of the evaluation set(3 enroll-1 auth utterances for each trial) and distribution of TC, TW, IC and IW trials in the evaluation set.

## 3. Methodology

### 3.1. Joint Factor Analysis

#### *3.1.1. Modeling*

Joint factor analysis can be regarded as modeling speaker and session variability in feature space by Gaussian Mixture Models(GMM's). [1, 5] It makes some basic assumptions regarding those variabilities stating a speaker and channel-dependent supervector of means M can be represented as

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

The channel supervector can be represented further as

$$\mathbf{c} = \mathbf{U}\mathbf{x} \tag{2}$$

where U is eigenchannel matrix for representing channel variations, and x is the channel vector matrix. [5]
And the speaker vector can be represented further by

$$\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} \tag{3}$$

where m is speaker and channel independent UBM(Universal Background Model) supervector, V is a low-rank matrix of eigenvoices and y is speaker factors vector. D is residual matrix for factors that can not be modeled by eigenvoices and z is the residual factors vector. [5]

MFCC are used as features in our system and the U, V, D matrices are estimated by Variational Bayes Inference on training set as: inferring first V assuming U and D is zero; then inferring U using estimated V and assuming D is zero; lastly inferring D using estimated V and U. [6, 7] In our system separate V,D,U's are estimated for target sentence and a closely related sentence. Then matrices that are trained on different data is concatenated in order to capture more variations. [8]

### 3.2. Scoring

Verification scheme starts with extracting x, y, z vectors for a given enrollment utterances by using MAP(maximum a priori) estimation from inferred V, U, D. Speaker and residual factor vectors y and z are used as feature vectors, and compared to vectors extracted from test utterance in terms of cosine distance. Similarity scores of y and z vectors averaged to have a fused feature vector.

ZT-norm score normalization is applied, by using gender dependent cohorts randomly selected from training speakers.

A basic GMM based gender classifier is trained on training set. Scoring is applied according to the detected gender among corresponding enrollments and ZT-norm statistics of that gender. ZT-norm also centers the scores from different JFA models, thus combining scores to same score range isn't a problem.

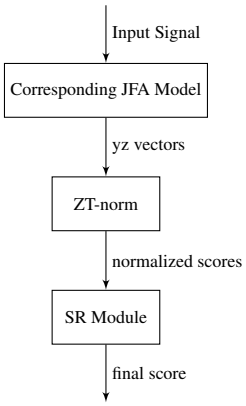
### 3.3. Speech Recognition Module

Speech recognition (SR) module is used for determining TW and IW trials. The phrases of the models are known therefore we only need to determine the phrase id of the evaluation utterances. When the phrase ids match the score of the biometric system is directly used. When the phrase ids do not match the output LR score of the biometric system is reduced.

Since there was a limit on the training data we only had librispeech for acoustic model training. We have used the pretrained Librispeech ASR Chain 1d model of Kaldi. So this is an english only model. Therefore we used english phonetic expansions of persian words.

For language modelling first a sentence model is generated. The words in all possible 10 sentences are combined and a recognition model is generated. Some of the utterances in the dataset did not produce matching results with the sentence model. So another model which is based on words is generated. After applying this word based recognition exact matches

Figure 1: *System Diagram*




to expected phrases are determined. When there is not an exact match to the sentence set, the recognized words are scored with sentences by matching word ratio. The utterance phrases are determined by combining results of sentence based SR model and word based SR model model. After determining that if the utterance phrase doesn't match target phrase, ZT-normalized likelihood score is decreased by a suitable amount in order to be rejected.

## 4. Results

Results are for system with and without SR module see the effect of target-wrong trials in the evaluation. Also eval score distributions with and without SR module can be seen in Fig.1.

Figure 2: *Eval Score Distributions*

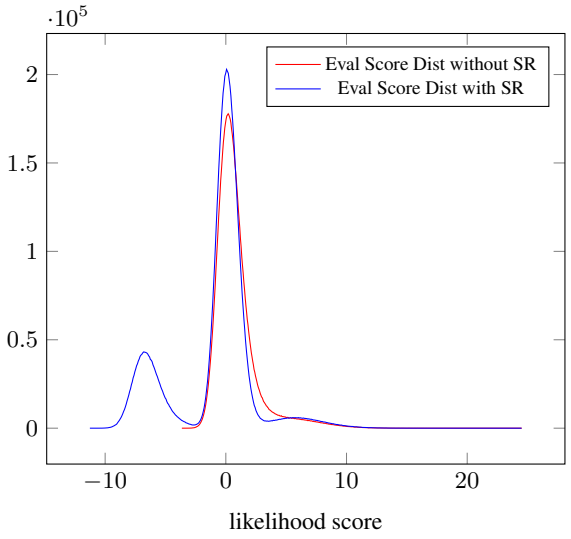



Table 1: RESULTS ON PROGRESS AND EVAL SET

| System | Set | EER | minDCF |
|---|---|---|---|
| System without SR Module | Progress | 4.616 | 0.236664 |
| System with SR Module | Progress | 2.922 | 0.102442 |
| System with SR Module | Eval | 2.958 | 0.102373 |

## 5. Conclusions

Joint Factor Analysis proved to be a robust baseline for text dependent speaker verification. Although it is affected by change in target utterance, without target detection of SR module and adjusting scores of tw trials; EER of 4.616 and minDCF of 0.24 may be still considered as acceptible.

## 6. References

[1] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," 2006.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.*, vol. 10, pp. 19–41, 2000.

[3] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[4] K. A. Zeinali, Hossein nad Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2020: the challenge evaluation plan." arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[5] P. Kenny, T. Stafylakis, M. J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," in *Odyssey*, 2014.

[6] T. Stafylakis, P. Kenny, M. J. Alam, and M. Kockmann, "Speaker and channel factors in text-dependent speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 65–78, 2016.

[7] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Comput. Speech Lang.*, vol. 22, pp. 17–38, 2008.

[8] N. Scheffer, R. Vogt, S. S. Kajarekar, and J. W. Pelecanos, "Combination strategies for a factor analysis phone-conditioned speaker verification system," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4053–4056, 2009.