Version 1.0

2010 NIST HASR1 Evaluation
Cogent Systems
Technical Response

Cogent Document # IG-EXT-OT-1255-0.00 (1)

Document Revision History

Version	Date	Author	Comment
1.0	06/04/2010	Cogent's R&D	Initial version

Contents

This page was intentionally left blank.

1 Introduction

The 2010 NIST Speaker Recognition Evaluation is part of an ongoing series of evaluations conducted by NIST providing an important contribution to the direction of research efforts and the calibration of technical capabilities. The 2010 evaluation also includes, for the first time, a Human Assisted Speaker Recognition (HASR) test. The first fifteen trials are called HASR1.

This document provides a brief description of the human-aided speaker recognition system used for Cogent Systems' (Cogent) submission to the 2010 NIST HASR1 evaluation.

Three methods are used for the HASR1 evaluation system, including (1) the traditional forensic method, which compares the formants, pitch and other parameters, (2) listener test, and (3) automatic speaker recognition. Due to the time/resource limitation, not all trials went through all three procedures. Automatic speaker recognition made the major contribution to the final recognition decisions.

2 ASR Sub-System

The ASR sub-system is a standard GMM-based system. Short time MFCC features, feature projection and session mismatch techniques are used. Score normalization is also applied.

The speaker model is derived by MAP adaptation from a universal background model (UBM). The channel of speaker model and UBM model and the T-Norm-models were all adapted by the test conversation.

2.1 Pre-process

The speech data is filtered by a high pass filter with the cut-off frequency of 180Hz. Then the data scale is normalized to the range of [-1, 1]. Next step is pitch based voice active detection, followed by data being emphasized with a factor of 0.97.

Figure 1.1 – Pre-process scheme

2.2 Feature extraction

Features used are short time cepstral coefficients (16 MFCCs), which are extracted over a 16ms window (hamming window applied) with a 8ms window shift, followed by feature warping with 3s window (375 frame). Then the feature vector is augmented with the first and the second derivatives giving a total 48 features. Finally, the features were projected by a HLDA matrix.

Figure 1.2 – Feature extraction scheme

2.3 Universal background model

The NIST2004 dataset was used for training the UBM model, including all 310 speakers' recordings. Single gender independent GMM with 2048 components were trained by EM algorithm. After that, the HLDA matrix was trained (no dimension reduction), and the UBM model was re-trained by the projected features and projected model with several EM iterations.

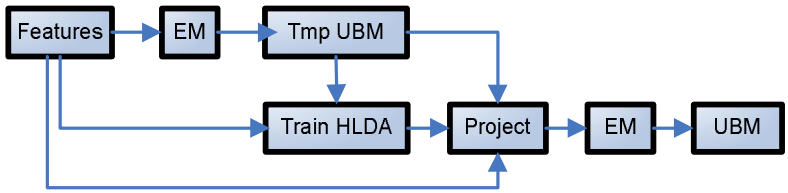

Figure 1.3 – UBM training scheme

2.4 Eigen-channel space

The eigen-channel space (U) was estimated by principal component analysis (PCA) on the NIST2004 dataset. All 310 speaker's voice data was used (2916 recordings, at least 2 recordings for each speaker). The channel number is 30.

2.5 T-Norm models

The T-Norm models (T) were selected from the NIST 2006 dataset of 100 males and 100 females.

2.6 Verification stage

Speaker's GMM model was derived from the UBM model with MAP adaptation. Target speaker GMM and the UBM were adapted for each trial using the eigen-channel adaptation. The adaptation was also applied to T-normalization.

For all speaker models (T) for each recording, the 5 highest-scoring UBM components were stored for each frame and used for computation of log-likelihood for speaker models.

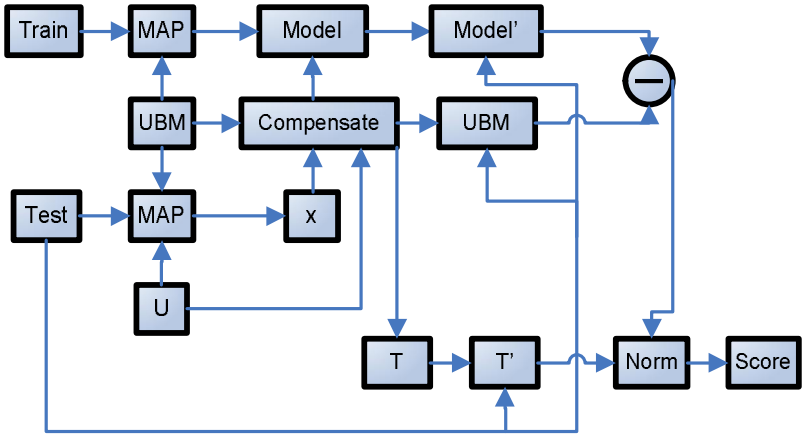

Figure 1.4 – Verification scheme

2.7 Result

Trail number	1	2	3	4	5	6	7	8
System score	2.01	0.58	-0.13	0.79	0.04	-0.21	1.31	0.16
System result	T	F	F	F	F	F	T	F
Really result	T	F	F	F	T	F	T	F
Error Count	8	13	8	8	8	11	11	7

Trail number	9	10	11	12	13	14	15
System score	0.46	0.96	0.46	0.57	-0.25	1.23	0.14
System result	F	T	F	F	F	T	F
Really result	F	T	F	F	F	T	T
Error Count	9	2	15	7	8	4	13

Figure 1.5 – Trail results

3 Human-Assisted methods

Two Human-Assisted speaker recognition methods were deployed. One is a listener test designed by our HASR team, and the other is the traditional forensic method. The speech was extracted from the .sph file, and the speaker of interest was specified in the NIST-provided information file.

3.1 Listener test

A group of 20 people participated in the listener test. The trail data was displayed to all participants and each participant gave a score, ranging from 0 to 1, to indicate his/her degree of certainty about whether the speaker in the model segment is present in the test segment on the channel of interest.. The scores from all listeners are used for a final likelihood score.

Due to time limitations, the listener test was only applied to some of the trials. All trials involving listener test returned correct results.

3.2 Forensic Method

Cogent performed the forensic analysis using the following procedure.

We first labeled the speech at the phoneme level, noting only the vowel phoneme. Then we checked the spectrogram and also listen to it for collecting the regions of interest (ROI). After that we extracted the parameters (formant, F1~F4) for comparison.

This method is very time-consuming so we only experimented with it and the results only contributed to the decision of one trial.

4 Summary

We consider the HASR evaluation very helpful that it encourages the efforts to bridge the gap between the automatic speaker recognition and the forensic speaker recognition. The process of building human aided speaker recognition system led to many good ideas and experiences. This is the first time for Cogent to participate in the Speaker Recognition Evaluation test and we value the experience.

In the future, we hope we can develop better strategies for combining the strength of ASR and human-assisted methods.