論文 / 著書情報
Article / Book Information

Title	Tokyo Tech Speaker Recognition
Authors	Sangeeta Biswas, Johan Rohdin, Koichi Shinoda
Citation	NIST SRE 2012, , ,
Pub. date	2012, 12

NIST SRE 2012 : TokyoTech Speaker Recognition

Sangeeta Biswas, Johan Rohdin, Koichi Shinoda

Department of Computer Science, Tokyo Institute of Technology, Japan

Introduction

- TokyoTech participated in the core condition
- Two systems were submitted:
  - Primary System : i-Vector + GPLDA
  - Alternate System : GMM-SVM
- Both of our systems performed terribly bad

System Configuration

- Frontend [HTK]
  - 15 PLP + log E + their Δ and ΔΔ (48 dimensions)
  - No feature warping
  - Energy based VAD for enrollment and test data, transcripts for UBM, T and development data
- UBM [HTK]
  - 512 Gaussian components
  - Diagonal co-variance matrices
  - 5-iterations of maximum-likelihood estimation
  - NIST SRE 2004
    - 75.1 hours of speech for male UBM
    - 110.5 hours speech for female UBM
- I-vector system [jfa-cookbook]
  - Dimension of i-Vector is 300
  - T matrix from NIST SRE 2004
    - 1337 speech files for male
    - 1740 speech files for female
    - 10 iterations for EM training
  - GPLDA [Gaussian PLDA package]
    - Same data as for T
    - 150 eigenvoices
    - i-Vectors were whitened and length normalized
- GMM-SVM system [LIBSVM]
  - Linear kernel
  - 4000 imposter files from NIST SRE 2004 data
  - NAP
    - 50 dimensions for the session subspace
    - Projection matrix trained using NIST SRE 2004 training data
- No score Normalization
- Score calibration [Bosaris toolkit]
  - Linear calibration
  - SRE06 as development data

Development Data

Dataset :

1conv4w-1conv4w task of the 2006 NIST SRE

- 813 speakers
- 2971 true trials and 30584 false trials in total

Results:

System	EER (%)	MDC
i-Vector	9.3	0.063
GMM-SVM	6.6	0.050

Results for NIST SRE 2012

System	EER (%)	MDC
i-Vector	40.5	0.095
GMM-SVM	49.1	0.100

Problem in i-Vector based System

Our analysis male SRE12 mic_phn showed three reasons of our bad performance

1. Inconsistencies in feature extraction

Automatic VAD was used for SRE12 but transcripts for SRE06, UBM and T data

Set	UBM + T	enroll + test	MDC	EER (%)
SRE06	Transcripts	VAD	0.091	19.6
SRE06	VAD	VAD	0.055	8.8
SRE12	Transcripts	VAD	0.100	42.3
SRE12	VAD	VAD	0.100	30.2

2. Weak T matrix

We used very little data for training T (and UBM)

- We added more data [SRE04, SRE05, Switchboard2 Phase1 (Swb2p1), Switchboard2 Phase2 (Swb2p2), Switchboard Cellular Part2 (SwbCellp2)]

Set	MDC	EER (%)
SRE06	0.054	7.6
SRE12	0.100	25.7

3. Different properties of SRE12 data set

- We compared the properties of the various data sets (SRE04, SRE05, Swb2p1, Swb2p2, SwbCellp2) in the following way
  1. For each set:
     A) Calculate the speaker dependent supervectors (By ML)
     B) Take the average of those supervectors to be the *representative* supervector of the set
  2. Calculate the cosine similarity between all the set representative supervectors
- Similarities between SRE12 and other sets are 0.60-0.64
- Similarities between other sets are 0.93-1.00
- Can these problems be solved by feature warping, noise cancellation etc.?

	Sre 04	Sre 05	Swb 2p1	Swb 2p2	Swb C. p2	Sre 06	Sre 12
Sre04	1.00	0.99	0.94	0.94	0.97	0.98	0.62
Sre05	0.99	1.00	0.95	0.95	0.98	0.98	0.62
Swb2p1	0.94	0.95	1.00	1.00	0.94	0.95	0.64
Swb2p2	0.94	0.95	1.00	1.00	0.93	0.95	0.64
SwbCellp2	0.97	0.98	0.94	0.93	1.00	0.97	0.60
Sre06	0.98	0.98	0.95	0.95	0.97	1.00	0.61
Sre12	0.62	0.62	0.64	0.64	0.60	0.61	1.00