HKPolyU System for NIST 2010 Speaker Recognition Evaluation

M.W. Mak

May 2010

Feature Extraction

	Specification/Method
Acoustic Features	• 12 MFCC + 12 ΔMFCC • 100 Hz frame rate • 25ms per frame
Feature Normalization	CMN followed by feature warping
Voice Activity Detection	Spectral subtraction followed by energy-based detection with cross talk removal

Speaker Modeling

	Specification/Method

UBMs	• 512 centers, diagonal covariance
	• Gender- and channel-dependent
	• Channel: mic or tel
	• K-means followed by EM
	• Telephone speech UBM:
	  • Data selected from NIST04+05+06
	• Mic speech UBM:
	  • Data selected from NIST05+06

GMM Speaker Models	• Gender- and channel-dependent MAP • Relevance factor = 16

GMM-SVM Speaker Models	• 12288-dim GMM-supervectors
	• 300 gender- and channel-dependent background speakers (one utterance per speaker) as impostor-class data
	• SVM penalty factor
	  • C=100 for speaker-class
	  • C=1 for impostor-class
	• Use the SVM Toolbox for Matlab provided by Anton Schwaighofer (Same performance as svmlight but much faster)

Channel Compensation

	Specification
NAP Corank	• Corank = 16 for telephone speech • Corank = 128 for microphone/interview speech

Training Data	• Telephone speech
	  • 457 male speakers from NIST04,05,06
	  • 861 female speakers from NIST04,05,06
	• Microphone speech
	  • 143 male speakers from NIST05,06,08
	  • 178 female speakers from NIST05,06,08
	• Each speaker has at least 8 utterances

NAP Matrices	• Gender-dependent • Speech-type dependent

Score Normalization (T-norm)

	Specification

Training Data	• Positive-class data:
	  • Telephone speech:
	  • 261 male utterances from NIST05
	  • 277 female utterances from NIST05
	  • Microphone speech
	  • 300 male speakers from NIST05,06
	  • 300 female speakers from NIST05,06
	• Impostor-class data:
	  • 300 gender- and channel-dependent background utterances

GMM-SVM T-norm Models	• Gender and speech-type dependent
	• SVM penalty factor
	  • C=100 for speaker-class
	  • C=1 for impostor-class

Score Fusion

NIST10 System 1:Fusion of 5 Systems

GSV+FSH+JSV+JFA+JSF

FSH: Fishervoice
GSV: GMM-SVM
JSV: JFA-SVM
JSF: JFA-SVM on channel factor

Score Fusion

NIST10 System 2: Selecting the Best System in NIST08

Gender	Condition	System
Male	phonecall_tel	FSH
Male	phonecall_mic	GSV
Male	interview	GSV
Female	phonecall_tel	FSH
Female	phonecall_mic	GSV
Female	interview	GSV

FSH: Fishervoice
GSV: GMM-SVM

Score Fusion

NIST10 System 3: Selecting the Best-3 Systems in NIST08

Gender	Condition	System
Male	phonecall_tel	GSV+FSH+JSV
Male	phonecall_mic	GSV+FSH+JSV
Male	interview	GSV+FSH+JSV
Female	phonecall_tel	GSV+JFA+FSH
Female	phonecall_mic	GSV+FSH+JSV
Female	interview	GSV+FSH+JSV

FSH: Fishervoice
GSV: GMM-SVM
JSV: JFA-SVM

Computation Time

Task	
CPU	CPU: Intel Core 2 Quad CPU Q9550 @ 2.83GHz
Operating System	Linux Fedora 11 (Kernel 2.6.27, 32 bits)
Memory	4G

Task	**CPU Time (sec)**
Training NAP matrix (phone speech)	711.0
Training NAP matrix (mic speech)	637.9

Computation Time

CPU: Intel Core 2 Quad CPU Q9550 @ 2.83GHz (Running-time on 1 core)

Task	CPU Time (sec) per utt.	% of Real-Time
VAD with Crosstalk Removal (8min)	103.0	21.5%
VAD w/o Crosstalk Removal (5min)	23.5	7.8%
VAD with Crosstalk Removal (3min)	19.7	10.9%
Feature Extraction with FW (8min)	34.7	7.2%
Feature Extraction with FW (5min)	29.5	9.8%
Feature Extraction with FW (3min)	12.4	6.9%
Creating GMM-Supervector (8min)	85.1	16.1%
Creating GMM-Supervector (5min)	68.1	22.7%
Creating GMM-Supervector (3min)	30.2	16.8%
Creating GMM-SVM	5.7	1.2%
GMM-SVM Scoring (8min)	8.1	1.7%
GMM-SVM Scoring (5min)	8.1	2.7%
GMM-SVM Scoring (3min)	8.1	4.5%

Computation Time

Enrollment time per 8min-utterance:

103+34.7+85.1+5.7 = 228.5 sec = 47.6% of real-time

Enrollment time per 5min-utterance:

23.5+29.5+68.1+5.7 = 126.8 sec = 42.3% of real-time

Enrollment time per 3min-utterance:

19.7+12.4+30.2+5.7 = 68.0 sec = 37.7% of real-time

Scoring time per 8min-utterance:

103+34.7+85.1+8.1 = 230.9 sec = 48.1% of real-time

Scoring time per 5min-utterance:

23.5+29.5+68.1+8.1 = 129.2 sec = 43.1% of real-time

Scoring time per 3min-utterance:

19.7+12.4+30.2+8.1 = 70.4 sec = 39.1% of real-time