CLIK (CMU+LIMSI+KIT) 2010 Speaker Recognition System

Qin Jin, Runxin Li, Qian Yang and Tanja Schultz
{qjin,lirx, tanja}@cs.cmu.edu; {qian, tanja}@ira.uka.de
Claude Barras, Viet-Anh Tran, Viet-Bac Le and Jean-Luc Rouas
{barras, tranviet, levb, rouas}@limsi.fr

General overview

CLIK is the result of a joint effort of speech processing teams from Carnegie Mellon University (Pittsburgh, USA), LIMSI-CNRS (Orsay, France) and KIT (Karlsruhe, Germany). This collaboration is carried out in the context of the French-funded QUAERO program on structuring and indexing of multimedia documents involving KIT and LIMSI [www.quaero.org].

Five sub systems are fused via linear regression: CMU-KIT contributes with two GMM/UBM+JFA sub-systems; LIMSI contributes with two GMM/UBM and one GSV-SVM sub-systems. Developments focused on the required core test, based on English telephone and microphone speech for 3 generic common conditions: interview-interview, interview-telephone and telephone-telephone.

Individual sub-systems

LIMSI

Only simple MFCC-GMM and GSV-SVM systems; LIMSI'08 [C]MLLR & prosodic systems (Ferras, 2007-2010, Leung, 2008) not presented due to technical and human constraints...

Sub-systems	Features	Models	Use of ASR	Intersession Variation Compensation
L1	MFCC	GMM	No	Factor analysis
L2	MFCC	GMM	No	Factor analysis
L3	Gaussian supervector of MFCC	SVM	No	NAP

MFCC-GMM: L1 & L2 differ only by the corpora used for training and normalization

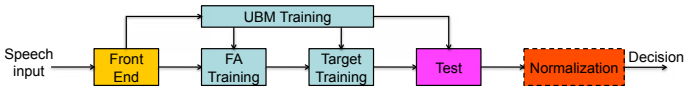
System	Features	Model		Gender-dependent Factor analysis (ALIZE)	ZT-norm
			Training data		
L1	• Voiced frames (pitch detection) • 15PLP + E + Δ + ΔΔ • Feature warping	• 256 Gaussians • MAP adaptation of gender-dependent UBMs	UBM trained from SRE04 tel	SRE06 tel	SRE06 tel
L2			UBM trained from SRE04 tel + SRE05 mic + SRE06 mic	SRE05 mic + SRE06 (tel+mic)	SRE05 + SRE06 (tel + mic)

GSV-SVM (L3)

System	Features	Model	Intersession variation compensation
L3	• Voiced frames (pitch detection) • 15PLP + E + Δ + ΔΔ • Feature warping • Feature mapping	• Gaussian mean supervector as feature - Gender-dependent 256 Gaussians - Variance normalization - Min-max normalization • Linear Kernel SVM classifier • SRE'04 telephone training data	NAP, 50 dimensions projected out, trained from SRE06 telephone data

CMU/KIT

CMU-KIT developed two sub-systems based on GMM/UBM+JFA, which differ in the front-end features. The system architecture is shown in the figure below.



Front-end Feature Extraction

- Sub-Sys C1: 20 MFCC + Δ (16ms winsize + 10ms shift)
- Sub-Sys C2: 13 MFCC + Δ (25ms winsize + 10ms shift)
- Cepstral mean subtraction and feature warping over 3s window (300 frms)
- Both sub-systems use the VAD provided by NIST for interview speech
- For telephone speech
  - Sub-Sys C1: bottom 30% of frames are excluded as nonspeech according to energy
  - Sub-Sys C2: 3 Gaussian classifiers based on the C0 feature

UBM Training

- gender-dependent UBMs trained using SRE06 training data, 1024 mixtures

Compensation with Joint Factor Analysis

- 300 Eigenvoice factors and 100 Eigenchannel factors
- 8 conversations utterances from SRE04, 05 and 06 training data for Eigenvoice training for both telephone and interview conditions
- Same data for Eigenchannel training for telephone
- SRE05 auxiliary microphone data for Eigenchannel training for interview

Normalization

- ztnorm showed improvement on dev data, not able to contribute to final submission due to resources limitations

Score-level System Fusion and Performance Analysis

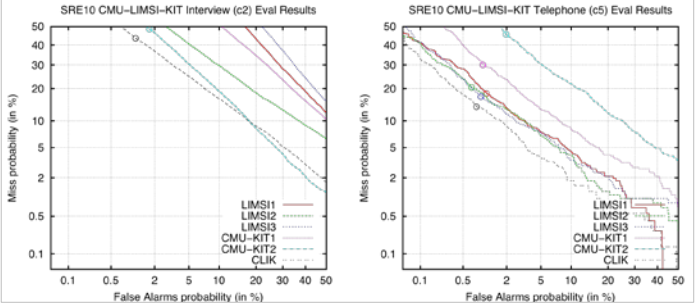
System fusion

- Linear logistic regression (FoCal toolkit, N. Brummer):
  SRE08 short2-short3 (core) trials for training fusion weights and decision threshold
- 3 separate fusion configurations for 3 common evaluation condition sets:
  interview-interview, interview-telephone and telephone-telephone.
- Fusion using SRE'08 cost, then shift scores using new SRE'10 cost

Performances with SRE'08 cost function

System	Interview-Interview				Interview-Telephone				Telephone-Telephone			
	SRE'08 (c1)		SRE'10 (c2)		SRE'08 (c4)		SRE'10 (c3)		SRE'08 (c7)		SRE'10 (c5)	
	MDC	%EER	MDC	%EER	MDC	%EER	MDC	%EER	MDC	%EER	MDC	%EER
L1	0.942	24.7	0.991	29.4	0.678	19.3	0.723	18.4	0.248	5.2	0.291	6.6
L2	**0.613**	14.2	0.738	19.1	0.650	16.6	0.542	13.2	0.258	5.5	0.271	**6.1**
L3	0.852	25.8	0.991	33.1	**0.510**	**12.7**	**0.482**	**11.0**	**0.219**	**5.0**	**0.262**	6.3
C1	0.687	13.7	0.904	26.8	0.779	22.8	0.662	15.7	0.279	6.1	0.399	8.9
C2	0.635	**12.7**	**0.658**	**13.9**	0.802	22.6	0.687	16.7	0.581	13.6	0.654	15.7
CLIK	**0.367**	**7.1**	**0.549**	**13.2**	**0.451**	**10.1**	**0.344**	**7.7**	**0.151**	**3.3**	**0.217**	**4.2**

- Large differences between sub-systems and conditions
- C2 (GMM/UBM+JFA) best on Interview-Interview
- Using microphone data useful (L2 > L1)
- L3 (GSV-SVM) best on telephone speech (no mic. speech in training)



Conclusions

- First collaborative effort between CMU, LIMSI and KIT (wiki, teleconfs…)
  - Simple MFCC-GMM and GSV-SVM systems
  - Specific sub-systems and fusion settings for interview data
- Interest in vocal effort condition, but not enough development data
- New MDC setting not well supported by SRE'08 data