SRI's NIST 2006 Speaker Recognition Evaluation System: Alternate Microphone Condition

Sachin Kajarekar, Luciana Ferrer, Martin Graciarena, Elizabeth Shriberg, Kemal Sönmez, Andreas Stolcke, Gokhan Tur, Anand Venkataraman

SRI International, Menlo Park, CA, USA

Collaborators:

ICSI group

Outline

- System overview
- Detailed results on SRE06
- Exploration before evaluation
- Summary and Future work

Submission Overview

- This is our first year of participation in this condition
  - Our main focus was the telephone-channel condition
- The system used here are same as the ones used for telephone-channel condition, except that the test waveforms were cleaned up using "wiener filtering"

Type	Features	Model	Trials Scored
Acoustic	MFCC	GMM	ALL
	MFCC	SVM	ALL
	Phone-loop MLLR 4 transforms	SVM	ALL
Stylistic	Word N-gram	SVM	English-only

Submission	Systems	Combiner
SRI_1 (primary) and SRI_2	SRI (4)	Neural Network(*)
SRI_3 (SRI+ICSI)	SRI (4) + ICSI (5)	

* SVM combiner training ran for really long time w/o converging … as if it is going in an infinite loop

More details

- English and non-English trials were combined separately using SRE05 altmic data
  - The output was normalized and thresholded (from SRE05) for each condition
  - The results are pooled into one submission
- This strategy worked fine except for 8-side non-English condition, SRE05 non-English data did not have any target trials
  - SRE05 altmic English data was used to train the combiner
- As it turns out, SRE05 and SRE06 data was similar and thresholds generalized very well

Results on SRE06

- English trials only (defined by v1 release) - 22160 for 1side and 3738 for 8sides
- Neural network combiner

1conv4w-1convmic results

Submission	#systems	%EER	DCF(x10)
SRI_1	4	6.26	0.23919
SRI_1, best result	4	6.26	0.23919
SRI+ICSI	9	5.83	0.23921
SRI+ICSI, best result	8 (w/o mllr)	5.79	0.20559

8conv4w-1convmic results

Submission	#systems	%EER	DCF(x10)
SRI_1	4	3.19	0.13157
SRI_1, best result	2 (cep_svm, word_ngram)	2.96	0.11274
SRI+ICSI	9	2.05	0.08934
SRI+ICSI, best result	6 (w/o gmm_cep, word_ngram, LC+WCCN)	2.05	0.08323

Results with SVM combiner

- It converged after around t+5 mins !!!
  - Relatively fewer systems and inclusion of C5 in training

1conv4w-1convmic results

Submission	Combiner	#systems	%EER	DCF(x10)
SRI_1	NN	4	6.26	0.23919
SRI_1	SVM	4	**5.95**	**0.23461**
SRI_3	NN	9	5.83	0.23921
SRI_3	SVM	9	**5.37**	**0.20986**

8conv4w-1convmic results

Submission	Combiner	#systems	%EER	DCF(x10)
SRI_1	NN	4	3.19	0.13157
SRI_1	SVM	4	**2.29**	**0.10974**
SRI_3	NN	9	2.05	0.08934
SRI_3	SVM	9	2.27	0.09773

Exploratory Experiments

- Before the evaluation, we tried following techniques on SRE05 altmic data
  - Wiener filtering
  - Feature transformation (using different microphones as classes)
  - Probabilistic optimal filtering (POF)
- Wiener filtering
  - Commonly used noise cancellation technique that estimates the noise from the silence region and uses the estimate to clean the speech
  - This is the only technique that gave improvement on SRE05 data

SRE05 altmic data

GMM cepstral system (w/o TNORM)	All 8 microphones		7 microphones (w/o C5)	
	%EER	DCF(x10)	%EER	DCF(x10)
w/o Wiener filtering	11.97	0.8016	10.5	0.7819
w/ Wiener filtering	11.34	0.7870	9.95	0.7687

Feature Transformation for GMM

- The idea is similar to Reynolds' 2003 paper on the same topic and the implementation is similar to the one used in our cepstral GMM system
- Algorithm
  - Create gender and microphone specific models from the background model
  - While training and testing, for each feature frame, chose 1-best Gaussian from the background model
  - For that Gaussian, find the gender-microphone model that gives the highest likelihood
  - Use the mean and the standard deviation of the most likely gender-microphone model to normalize the features
- Implementation
  - Gender-microphone models were trained using SRE04 altmic devset
    - On top of the gender-handset normalized features
    - Alongside with gender-handset models
  - Transform applied on SRE05
- Results showed marginal or no improvement in performance

Probabilistic Optimum Filtering (POF)

- Motivation: significant WER reductions in noisy and multiple microphones speech recognition
- Algorithm (Neumeyer & Weintraub, ICASSP '94)
  - POF mapping is piecewise linear transformation of the mismatched (noisy) feature space into the matched (clean) feature space
  - It requires stereo data: time aligned waveforms from the matched and mismatched cases
  - A VQ partition of the noisy feature space is first computed
  - Using the MMSE criterion a transformation is trained for each VQ region
  - In testing the clean feature estimate is computed by a weighted average of the region-specific clean feature estimates
- Implementation
  - POF trained using SRE04 altmic devset
- Results show no improvement in performance.

Issues with SRE04 and SRE05 altmic data

- There is a big mismatch between the two datasets
  - Different collection sites? different position of mics?
- Performance for some microphones is not much worse than the performance on clean condition
  - %EER is comparable
  - DCF is almost twice as bad

Microphone detector trained on SRE04 data

SRE05 MicType	% Detection
C1	21.57
C2	26.82
C3	86.25
C4	63.45
C5	84.54
C6	39.06
C7	44.89
C8	41.98

SRE05 Cepstral GMM system (w/o TNORM)

Chan	Teleph	C1	C2	C3	C4	C5	C6	C7	C8
%EER	8.18	7.69	7.31	12.69	10.3	21.92	10.74	10.74	7.72
DCF	0.3115	0.6934	0.6754	0.7848	0.7972	0.8682	0.7493	0.8220	0.7231

Summary and Future Work

- This was our first stab at the altmic condition. The submission used a subset of systems developed for the telephone condition.
  - SRI primary submission used 3 cepstral system and 1 stylistic feature based system
  - SRI secondary submission used SRI + ICSI systems
  - The performance is very competitive
- We explored three techniques for noise robustness
  - Wiener filtering – most successful
  - Feature transformation & POF – needs more work
- SRE04 altmic development data is very different from SRE05 altmic data
  - This hindered any efforts to investigate more sophisticated techniques
- Future work
  - Use all the SRI stylistic feature based systems on this data
  - Use SRE05 data for feature transformation and POF
  - SNR dependent processing