Language mismatch in speaker recognition system

student： Askar

Weekly meeting report

2014-12-29

GROUPING
Center for Speech and Language Technologies

Outline

1. Introduction
2. Feature level solution
3. Model level solution
4. Score level solution
5. references

1.Introduction

- Speaker recognition: recognize the identity of a speaker from speech.
- Categorization
  - verification and identification
  - text independent and text dependent
  - mono-lingual, cross-lingual and multi-lingual

1.Introduction

- Mono-lingual: the language of training and testing is the same
- Cross-lingual: speaker model is trained in one language and tested with a speech in another language
- Multi-lingual: training is done in one language and tested with a speech of multiple language.

1.Introduction

- Language mismatch between training and testing data leads to significant performance degradation

Table 1

UBM-Lang	GMM-Lang	Testing-Lang	EER (%)
Chinese	Chinese	Chinese	2.64
Chinese	Chinese	Uyghur	14.80

- In this report, we give overall introduction to the current research works of cross-lingual and multi-lingual speaker recognition

2.Feature level solution

1) Vocal source features for bilingual speaker identification

--*JiangLin Wang, Michael T. Johnson, ChinaSip, 2013*

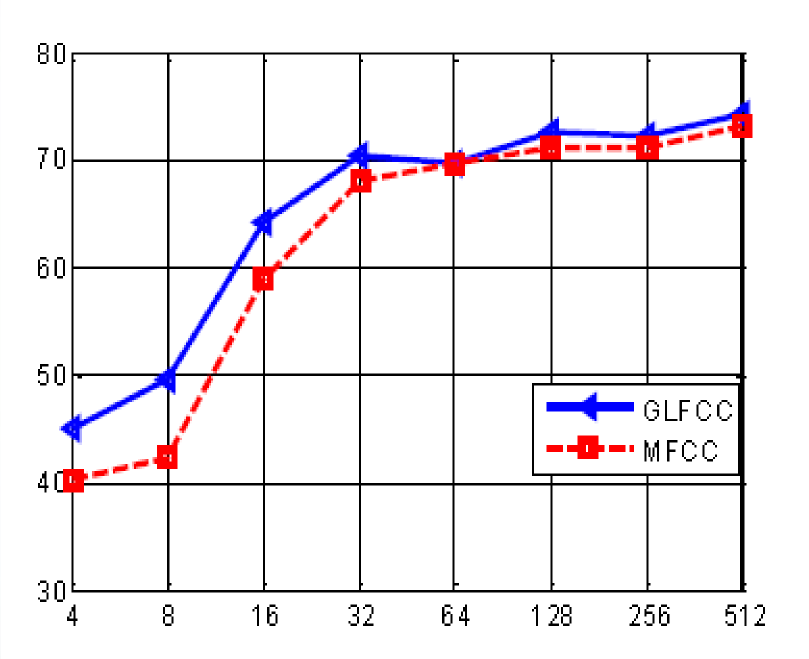
- Authors captured speaker-specific characteristics from their vocal excitation patterns using:
  - RPCC: Residual Phase Cepstrum Coefficients
  - GLFCC: Glottal Flow Cepstrum Coefficients

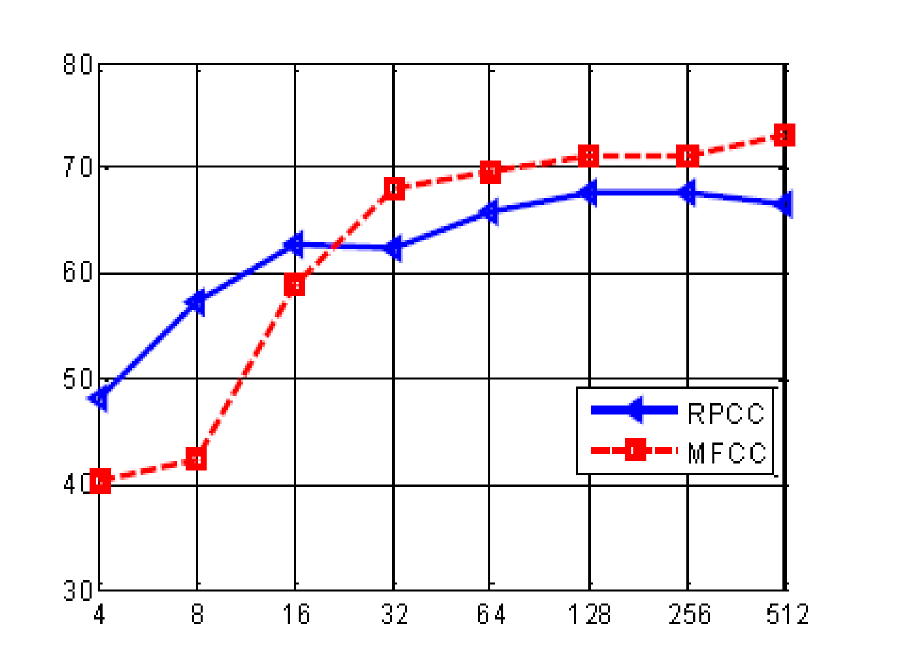
2.Feature level solution

- Data: speech of twenty-four bilingual speakers extracted from 2004 NIST SRE corpus.
- Considered Languages: Arabic, Mandarin, Russian and Spanish.
- UBM: trained using data from all twenty-four non-English speakers.
- GMM: adapted from UBM using individual English speech samples
- Identification: performed using alternative language speech samples.
- Baseline features: MFCC

2.Feature level solution

- Accuracy with increasing number of mixtures





2.Feature level solution

- Accuracy of individual features

Table 2

Individual Feature	Accuracy (%)
MFCC	71.2
GLFCC	72.3
RPCC	67.7

- GLFCC has the highest accuracy

- RPCC gives the highest accuracy with the small mixture number.

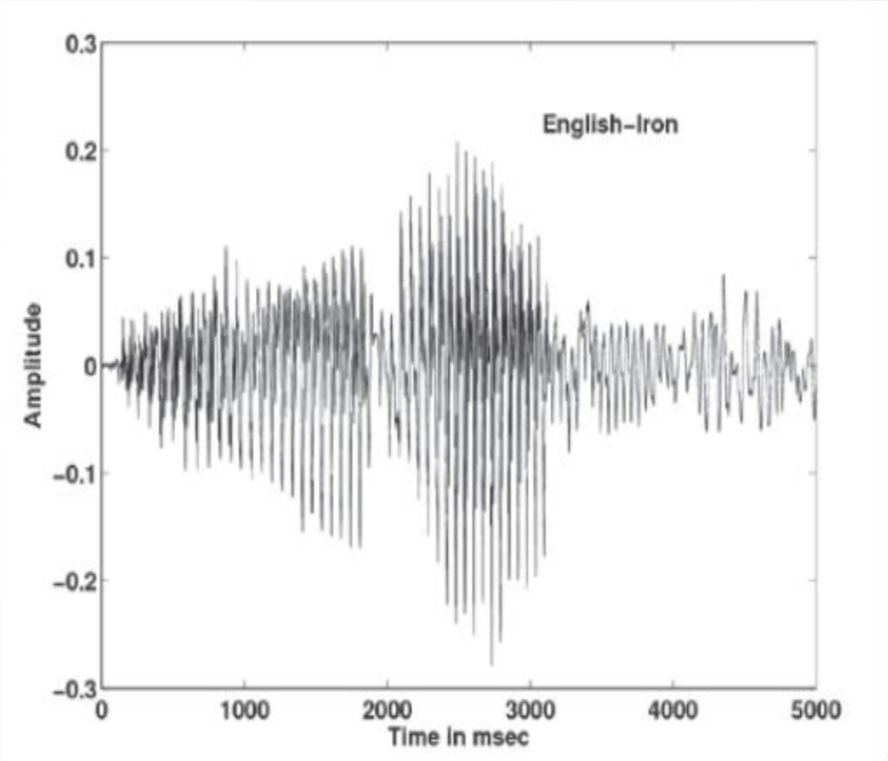
2.Feature level solution

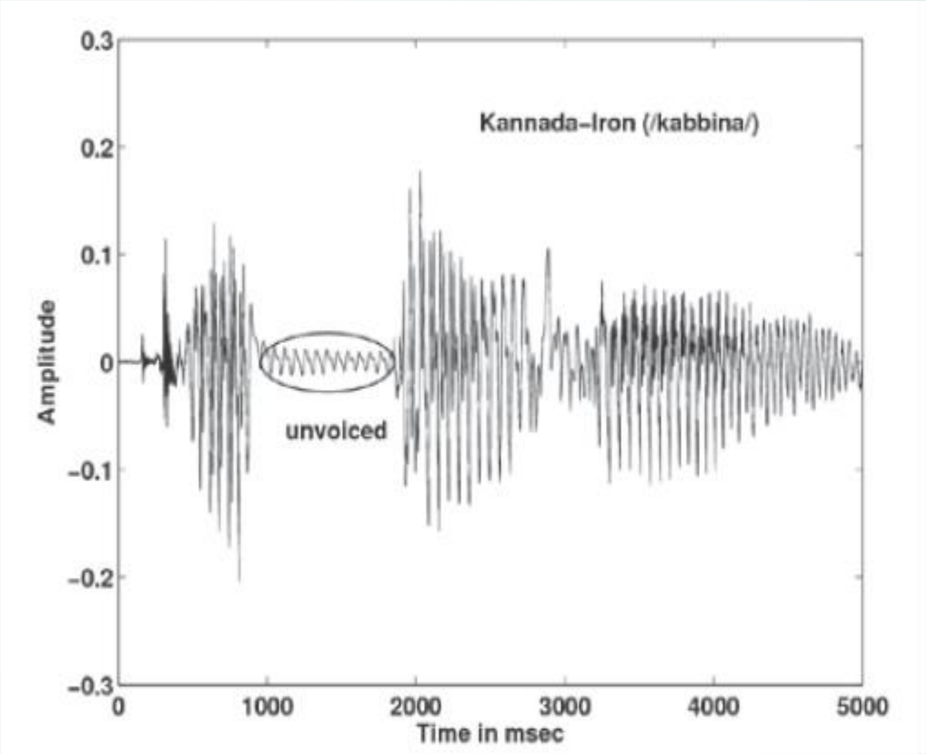
2) Kannada Language Parameters for Speaker Identification

--Nagaraja B.G. , I.*J. Image, Graphics and Signal Processing*, 2013

- Feature: MFCC feature
- Considered languages: English, Hindi and Kannada (regional language)
- The speaker utters a word in English, there is no much pause in the speech signal, but when he/she pronounces the corresponding word in Kannada there is a long pause in the speech signal

2.Feature level solution





2.Feature level solution

- The presence of ottakshara (CCV akshara like, /gga/ in /agga), arka (refers to a specific /r/ in consonant clusters) and anukaranavyayagalu (/julujulu/) leads to long pause and hence less number of energy frames in Kannada words.
- In order to alleviate this problem, a new database was created using the same speakers in Kannada language where words which are free from ottakshara, arka and anukaranavyayagalu

2.Feature level solution

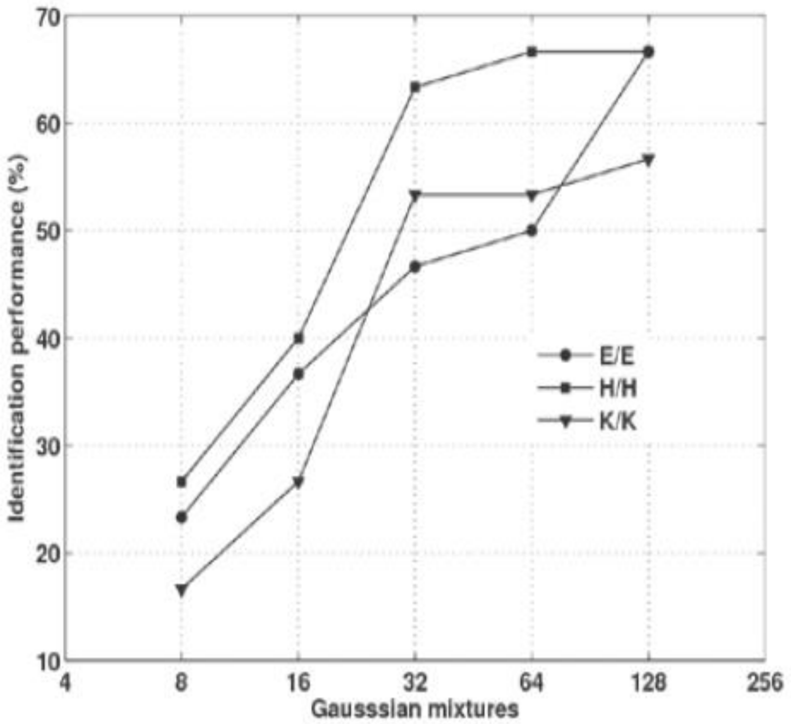


Figure 3. Monolingual speaker identification performance.

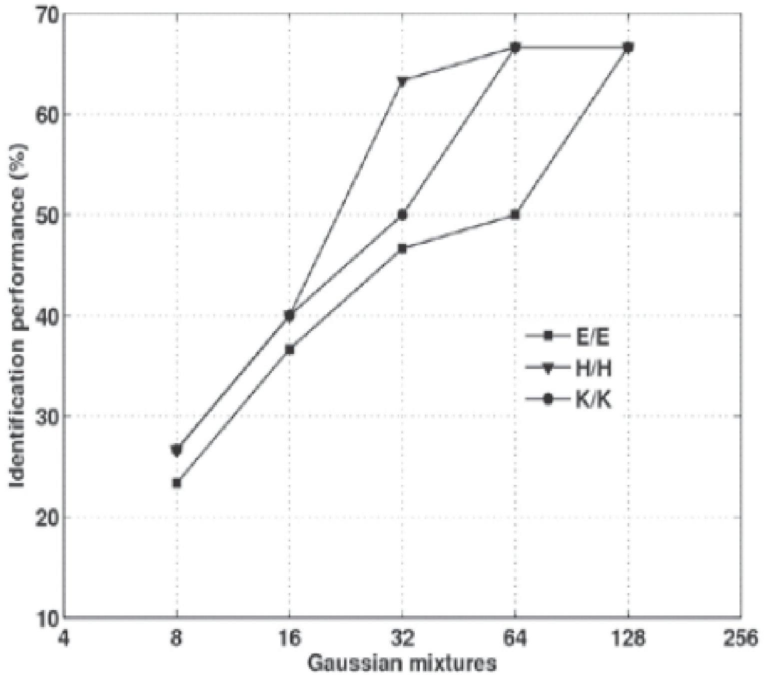


Figure 12. Monolingual speaker identification performance after considering the language parameters for Kannada.

2.Feature level solution

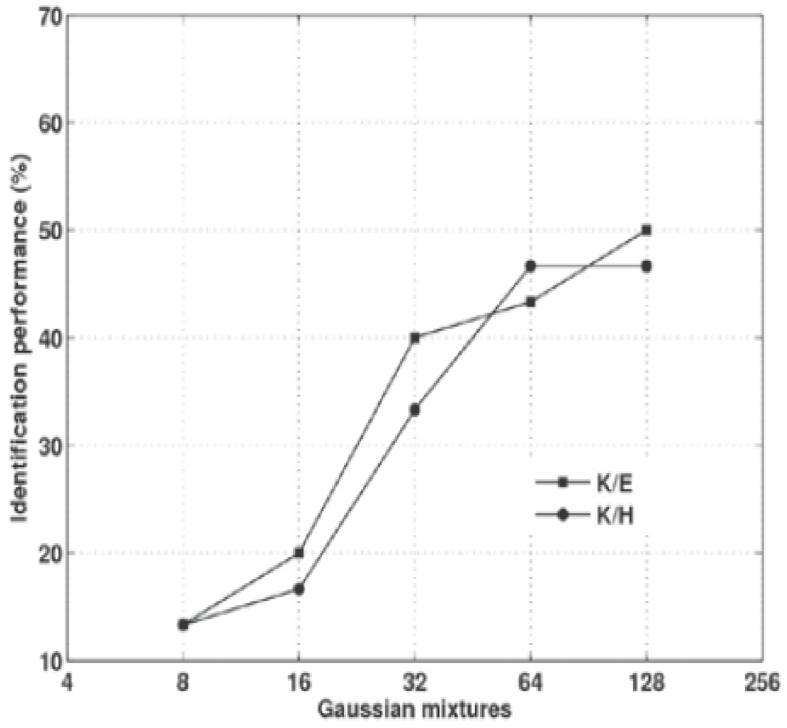


Figure 6. Crosslingual speaker identification performance (K/E and K/H).

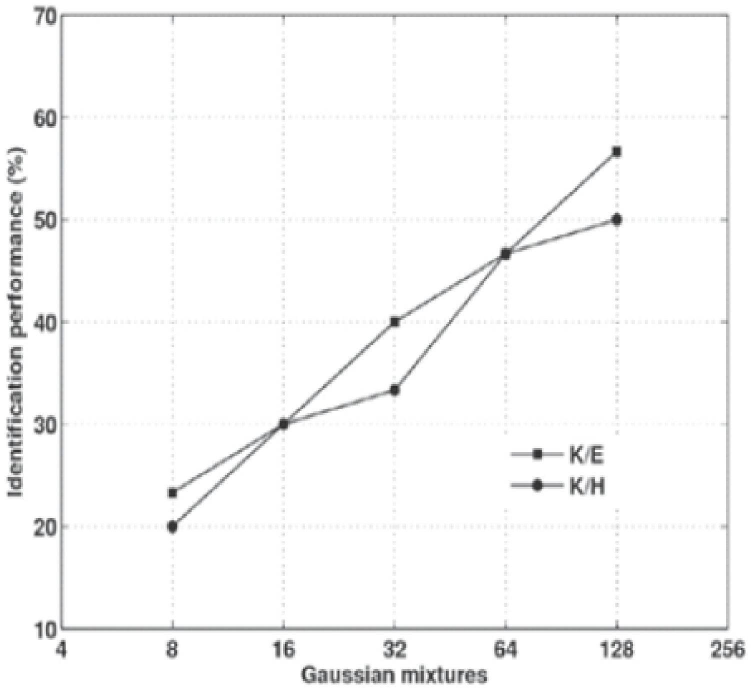


Figure 13. Crosslingual speaker identification performance (K/E and K/H) after considering the language parameters for Kannada.

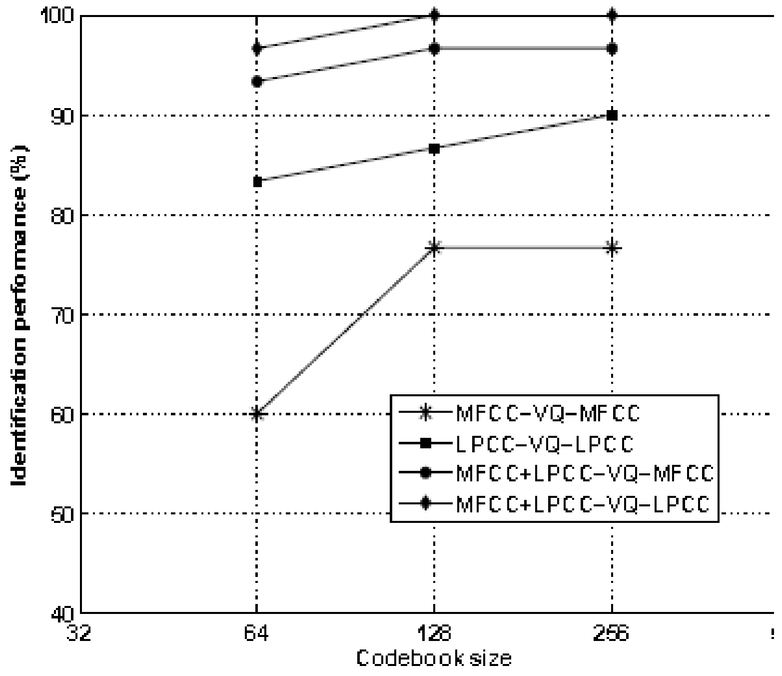
2.Feature level solution

3) Combination of Features for Multilingual Speaker Identification with the Constraint of Limited Data

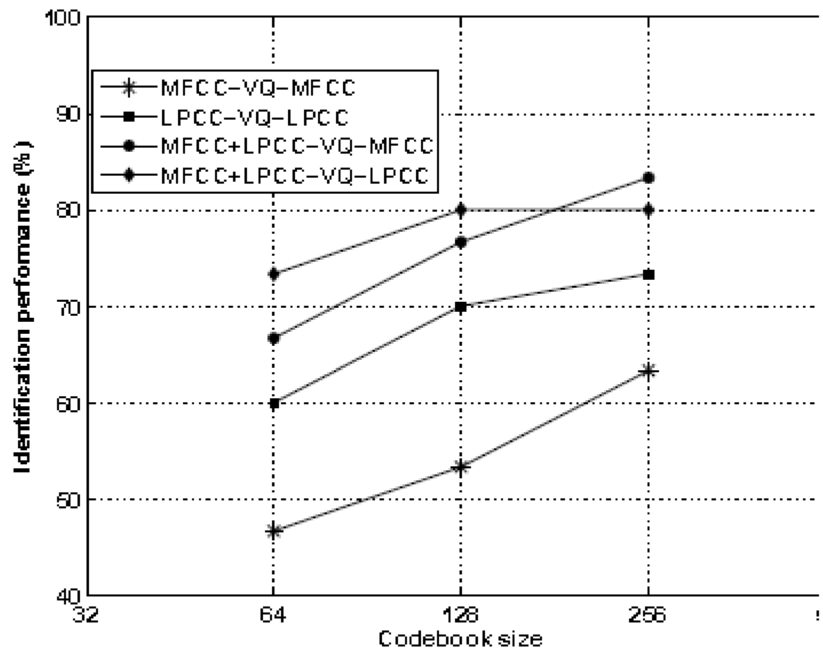
--Nagaraja B.G., I.J. of Computer Applications,2013

- Feature: combined features of MFCC and LPCC
- Considered language: English, Hindi and Kannada (regional language)
- Data: set of 30 speakers

2.Feature level solution



(a) E/E



(a) E/H

2.Feature level solution

Results of combined features of MFCC and LPCC

Speakers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	No
MFCC	√	√	X	X	X	X	X	√	√	√	√	X	X	X	X	√	X	√	√	√	X	√	√	√	√	√	√	X	√	√	18
LPCC	√	√	√	√	X	X	X	√	√	√	√	√	√	X	X	√	X	√	√	X	X	√	√	√	√	√	√	X	√	X	20
Combined	√	√	√	√	X	X	√	√	√	√	√	X	√	X	X	√	√	√	√	X	X	√	√	√	√	√	√	X	√	√	22

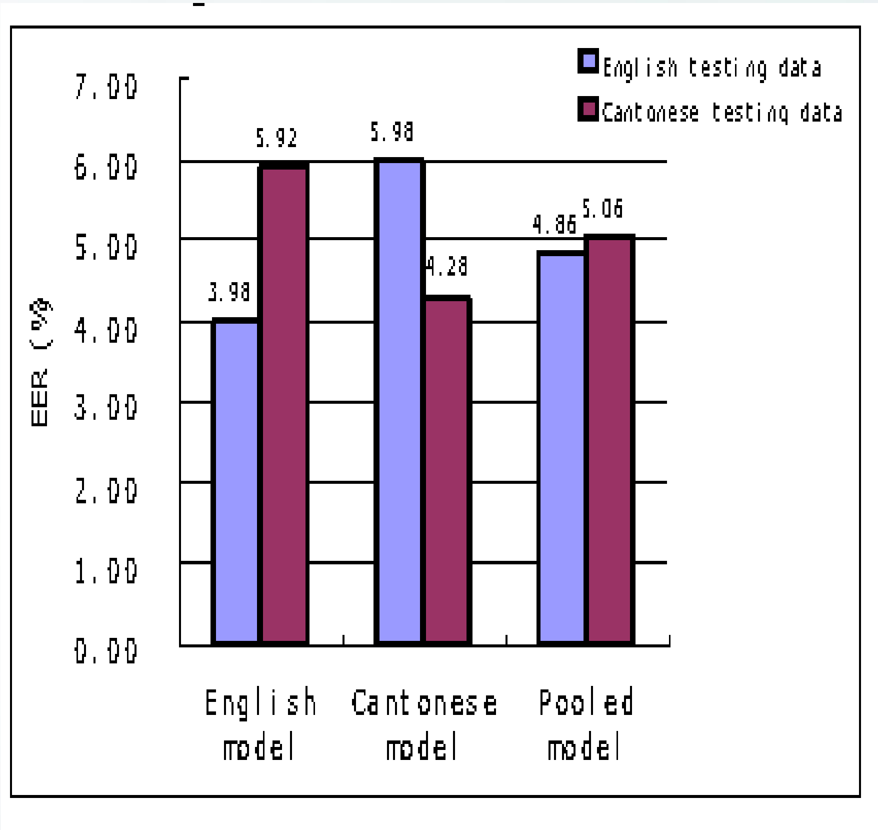
3.Model level solution

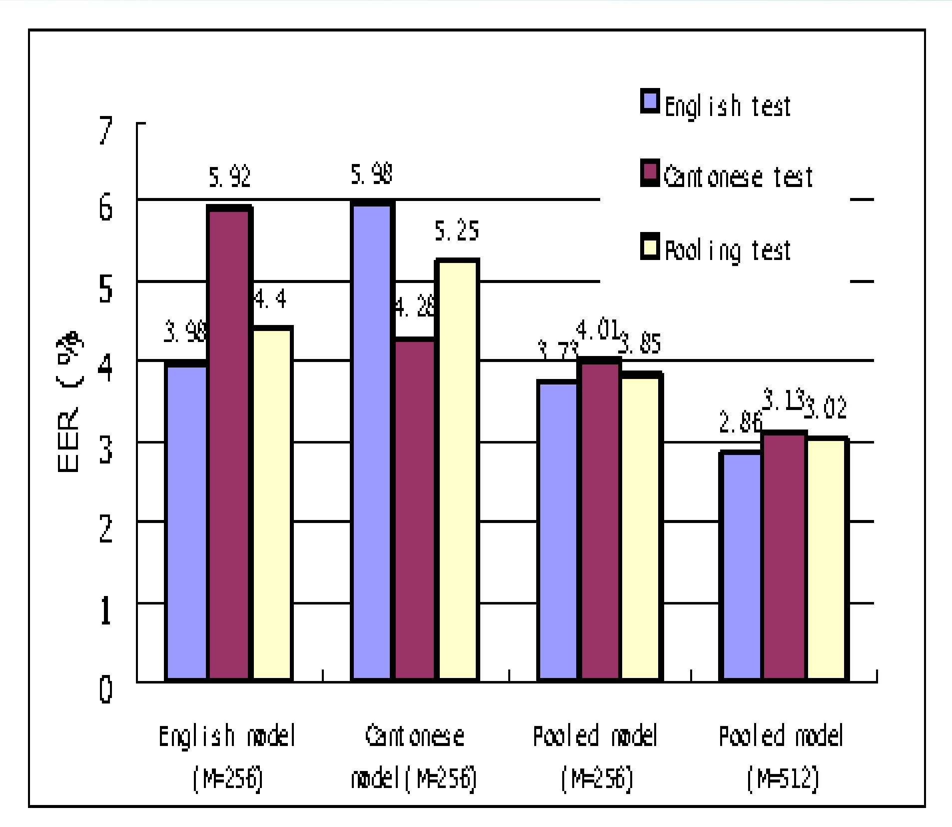
1) ENGLISH-CHINESE BILINGUAL TEXT-INDEPENDENT SPEAKER VERIFICATION

---Bin Ma and Helen Meng, ICASSP 2004

- Considered languages: English and Cantonese
- Data: self designed and collected CUHK bilingual speech corpus including prompts for commands and questions of personalized information
- Model: GMM trained with utterances from both languages.

3.Model level solution





3.Model level solution

2) THE EFFECT OF LANGUAGE FACTORS FOR ROBUST SPEAKER RECOGNITION

---Liang Lu, ICASSP 2009

- Considered languages: 18 languages including English
- Data: The Oregon Graduate Institute (OGI) multi-language corpus
  2004 and 2008 NIST SRE data
- Model: Extend JFA model with language factors.

3.Model level solution

- Language factor was enrolled based on the conventional joint factor analysis.
- Extend JFA model with language factors

$$M = m' + Bg + Vy + Dz + Ux$$

  - M : Speaker's GMM mean super vector
  - m' : speaker and language-independent supervector,
  - B : low-rank rectangular transformation matrix
  - g : language factors.
  - BB*: language subspace.

3.Model level solution

- Language subspace estimation

  a) remove speaker and session attribute of the multi-language data

  b) U=0, V=0 and D=0: assuming the speaker factors be averaged out because of the sufficient amount of data of each language.

  c) Randomly initialize B

  d) Calculate P($g(l)|x(l)$): Gaussian mean $\xi(l)^{-1}B\Sigma^{-1}\tilde{F}(l)$ and variance $\xi(l)^{-1} = I + B^{*}\Sigma^{-1}N(l)B$

  e) B is re-estimated via EM iteration

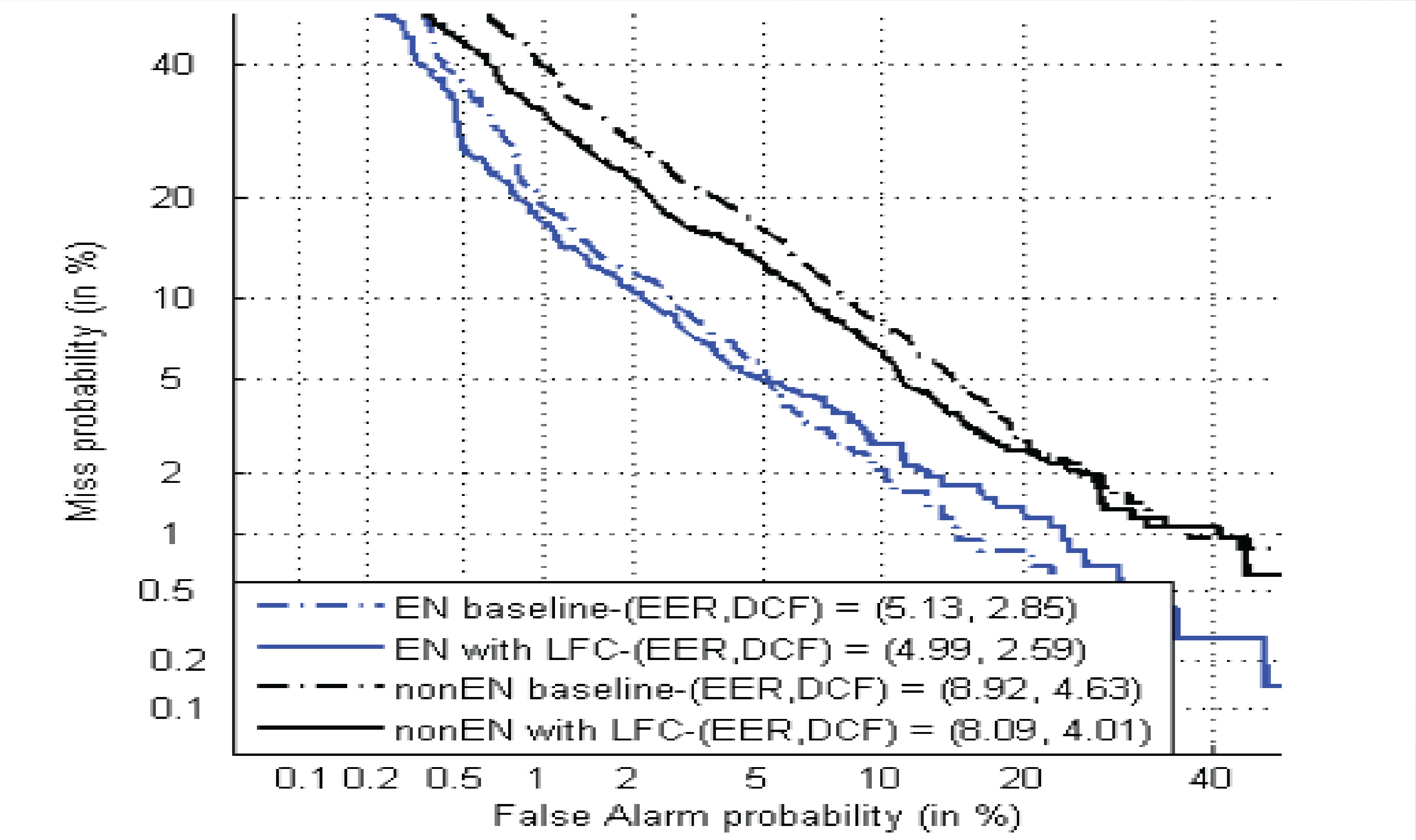
$$\prod_{l} \max_{g} P_{HMM}\left(\chi(l) \middle| m' + Bg, \Sigma\right)$$

3.Model level solution

- Language factor compensation
  - Training phase: the language factors of training utterances were removed from the models
  - Testing phase: compensation was performed in the model level, namely:

$$llr(X_{utt}, M_{tar}) = \frac{1}{T}\log\left(\frac{p(X_{utt}|M_{tar} + Bg_h)}{p(X_{utt}|M_{UBM} + Bg_h)}\right)$$

3.Model level solution



4.Score level solution

1) THE EFFECT OF LANGUAGE FACTORS FOR ROBUST SPEAKER RECOGNITION

---Liang Lu, ICASSP 2009

- Considered languages: 18 languages including English
- Data: The Oregon Graduate Institute (OGI) multilanguage corpus
  2004 and 2008 NIST SRE data
- Model: Extend JFA model with language factors.

4.Score level solution

- Score level fusion was made as follows:

$$s\left(X_{utt}\right) = \alpha \cdot llr_{ec}\left(X_{utt}\right) + \left(1-\alpha\right) llr_{gfc}\left(X_{utt}\right)$$

  - $\alpha \subset [0,1]$: weight parameter

4.Score level solution

Systems	English trails		non-English trails	
	EER	DCF	EER	DCF
Baseline	7.84%	.372	11.42%	.566
LFC only	7.11%	.328	9.8%	.417
eigenchannels only	5.03%	.223	11.19%	.412
Combination in model level	5.13%	.226	11.19%	.408
Combination in score level	**5.13%**	**.218**	**9.04%**	**.374**

4.Score level solution

2) LANGUAGE NORMALIZATION FOR BILINGUAL SPEAKER RECOGNITION SYSTEMS

---Murat Akbacak, John H.L. Hansen, ICASSP, 2007

- Considered languages: English and Spain
- Data: Miami Corpus
- Model: GMM

4.Score level solution

- Baseline system (B2): merges language-dependent systems' outputs via score fusion

$$\Lambda^{+} = \underset{1 \le n \le N}{argmax} \left[p(O|\Lambda_{n,Eng})\, w_{Eng} + p(O|\Lambda_{n,Spn})\, w_{Spn} \right]$$

  - w_{Eng}, w_{Spn}: fusion weights, optimized using development set

4.Score level solution

a) Normalization at the utterance level: LID scores corresponding to each language are used as fusion weights.

$$\Lambda^* = \underset{1 \le n \le N}{\operatorname{argmax}} \left[p(O|\Lambda_{n,Eng})\, p(Eng|O) + p(O|\Lambda_{n,Spn})\, p(Spn|O) \right]$$

- the probability of the event that the utterance is spoken in language L is used to weight the likelihood score coming from language dependent speaker recognition system

4.Score level solution

b) Normalization at the segment level: language-dependent speaker recognition system outputs are merged at the segment level.

$$S(n) = \sum_{i=1}^{M} p(O_i|\Lambda_{n,Eng})\, w_{i,Eng} + p(O_i|\Lambda_{n,Spn})\, w_{i,Spn}$$

- M: represents the number of segments

4.Score level solution

➢ segments corresponding to phones existing in both English and Spanish acoustic spaces are weighted more

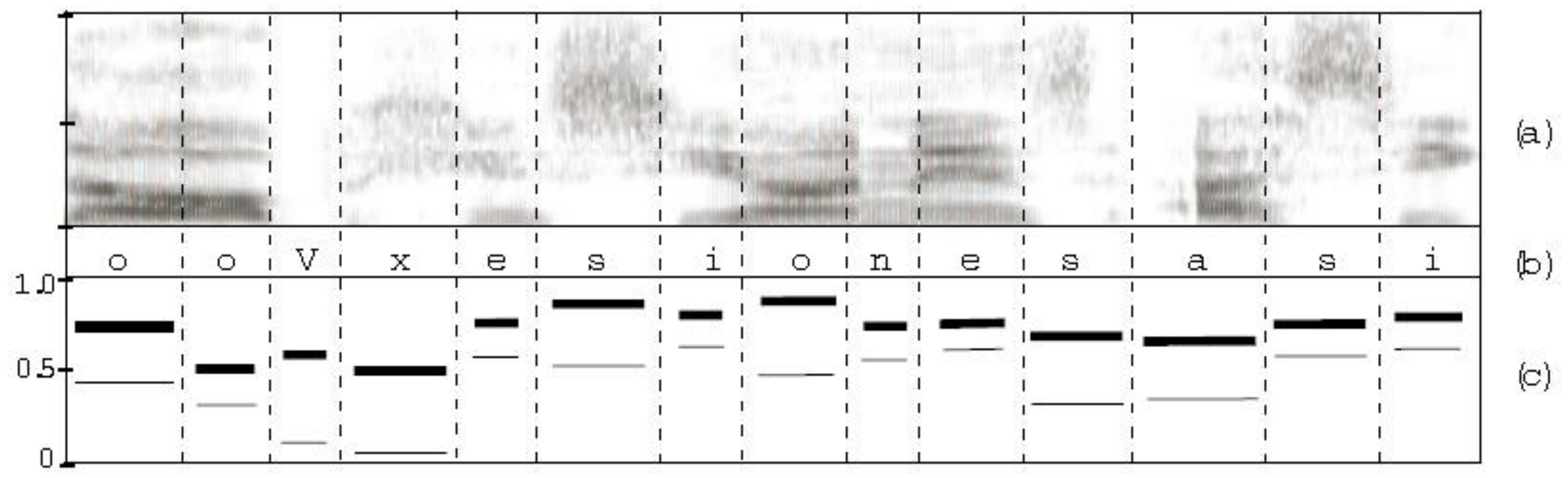


Fig. 2. Example to segment-based normalization for a Spanish utterance with its spectrogram (a), phonetic transcription (b), and normalization weights (c). Thick and thin lines correspond to $w_{i,Spn}$ and $w_{i,Eng}$ values respectively.

4.Score level solution

- Experimental result:

Exp.	Train	Test	B2	LID-norm	PR-norm
3	Spn	Eng	83.49%	83.49%	84.82%
4	Eng	Spn	70.31%	70.31%	74.31%
5	Eng + Spn	Eng	81.22%	82.13%	83.21%
6	Eng + Spn	Spn	80.05%	81.32%	82.37%

5. References

[1] Wang, Jianglin ; Johnson, Michael T, “Vocal source features for bilingual speaker identification”, ChinaSIP, pp.170-173, 2013

[2] Nagaraja B.G. ; H.S. Jayanna, “Kannada Language Parameters for Speaker Identification with The Constraint of Limited Data”, International Journal of Image, Graphics and Signal Processing, Vol 5, Iss 9, Pp 14-20 (2013), 2013

[3] B.g., Nagaraja ; S. Jayanna, H., “Combination of Features for Multilingual Speaker Identification with the Constraint of Limited Data ”, International Journal of Computer Applications, 2013, Vol.70(6), pp.1-6

[4] Bin Ma, Bin Ma ; Meng, H., “English-Chinese bilingual text-independent speaker verification ”, 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2004, Vol.5, pp.V-293-6

5. References

[5] Lu, Liang ; Dong, Yuan ; Liu, Jiqing ; Dong, Yuan ; Zhao, Xianyu ; Wang, Haila, “The effect of language factors for robust speaker recognition”, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2009, pp.4217-4220

[6] Akbacak, Murat ; Hansen, John H.L., “Language normalization for Bilingual Speaker Recognition systems ”, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2007, Vol.4, pp.IV257-IV260

[7] Auckenthaler, R. ; Carey, M.J. ; Mason, J.S.D., “Language dependency in text-independent speaker verification ”, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2001, Vol.1, pp.441-444

5. References

[8] Sarkar, Sourjya ; Rao, K. Sreenivasa ; Nandi, Dipanjan ; Kumar, S. B. Sunil, "Multilingual speaker recognition on Indian languages", 2013 Annual IEEE India Conference, INDICON 2013, 2013

[9] Haris, B.C. ; Pradhan, G. ; Misra, A. ; Shukla, S. ; Sinha, R. ; Prasanna, S.R.M, "Multi-variability speech database for robust speaker recognition", 2011 National Conference on Communications, NCC 2011, 2011

[10] Lamel, L.F. ; Gauvain, J.-L., "Cross-lingual experiments with phone recognition ", 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing 1993, Vol.2, pp.507-510

5. References

[11] Nagaraja, B.G. ; Jayanna, H.S., “Mono and cross lingual speaker identification with the constraint of limited data”, International Conference on Pattern Recognition, Informatics and Medical Engineering, PRIME 2012, 2012, pp.439-443

[12] Jin, Qin ; Schultz, Tanja ; Waibel, Alex, “Speaker identification using multilingual phone strings ”, Acoustics, Speech, and Signal Processing, May 2002, Vol.1, pp.I-145-I-148

[13] Goggin, Judith P. ; Simental, Liza R. ; Thompson, Charles P. ; Strube, Gerhard, “The role of language familiarity in voice identification”, Memory & Cognition, September 1991, Vol.19(5), pp.448-458

5. References

[14] Kohler, M.A. ; Andrews, W.D. ; Campbell, J.P. ; Herndndez-Cordero, J., "Phonetic speaker recognition ", Conference Record of Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, Nov. 2001, Vol.2, pp.1557-1561

[15] Wang, Jianglin ; Ji, An ; Johnson, Michael T., "Features for phoneme independent speaker identification", ICALIP 2012 - 2012 International Conference on Audio, Language and Image Processing, Proceedings, 2012, pp.1141-1145

[16] Künzel, Hermann J., "Automatic speaker recognition with cross-language speech material ", International Journal of Speech Language and the Law, 2013, Vol.20(1)

[17] Utpal Bhattacharjee ; Kshirod Sarmah, "GMM-UBM Based Speaker Verification in Multilingual Environments ", International Journal of Computer Science Issues, 2012, Vol.9(6), p.373

5. References

[18] Ranjan, Rajesh ; Singh, Sanjay Kumar ; Shukla, Anupam ; Tiwari, Ritu, “Text-dependent multilingual speaker identification for Indian languages using Artificial Neural Network”, Proceedings - 3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010, 2010, pp.632-635

[19] Pandey, B. ; Ranjan, A. ; Kumar, R. ; Shukla, A., “Multilingual speaker recognition using ANFIS”, Signal Processing Systems, July 2010, Vol.3, pp.V3-714-V3-718

[20] Sundaradhas Selva Nidhyananthan ; Ramapackiam Shantha Selva Kumari, “A FRAMEWORK FOR MULTILINGUAL TEXT- INDEPENDENT SPEAKER IDENTIFICATION SYSTEM”, Journal of Computer Science, Vol 10, Iss 1, Pp 178-189 (2014), 2014

5. References

[21] Agrawal, Prateek ; Shukla, Anupam ; Tiwari, Ritu, "Multi lingual speaker recognition using artificial neural network", Advances in Intelligent and Soft Computing, 2009, Vol.61, pp.1-9

[22] Prateek Agrawal ; Harjeet Kaur ; Gurpreet Kaur, "Multi Lingual Speaker Identification on Foreign Languages using Artificial Neural Network ", International Journal of Computer Applications, 2012, Vol.57(13), p.36

[23] Kenny, Patrick ; Boulianne, Gilles ; Ouellet, Pierre ; Dumouchel, Pierre, "Joint factor analysis versus eigenchannels in speaker recognition ", IEEE Transactions on Audio, Speech and Language Processing, May 2007, Vol.15(4), pp.1435-1447

[24] Kenny, Patrick ; Ouellet, Pierre ; Dehak, Najim ; Gupta, Vishwa ; Dumouchel, Pierre ; Kenny, Patrick ; Gupta, Vishwa ; Dumouchel, Pierre

5. References

[24] Kenny, Patrick ; Ouellet, Pierre ; Dehak, Najim ; Gupta, Vishwa ; Dumouchel, Pierre ; Kenny, Patrick ; Gupta, Vishwa ; Dumouchel, Pierre, “A study of interspeaker variability in speaker verification ”, IEEE Transactions on Audio, Speech and Language Processing, 2008, Vol.16(5), pp.980-988

[25] Dehak, Najim ; Kenny, Patrick J. ; Dumouchel, Pierre ; Ouellet, Pierre ; Dehak, Réda ; Dumouchel, Pierre, “Front-end factor analysis for speaker verification ”, IEEE Transactions on Audio, Speech and Language Processing, 2011, Vol.19(4), pp.788-798

[26] Abhinav Misra; John H. L. Hansen, “Spoken Language Mismatch in Speaker Verification: An investigation with NIST-SRE and CRSS Bi-Ling Corpora”, Spoken Language Technology Workshop, 2014

Thank You !

GROUPING

Center for Speech and Language Technologies