Score Fusion and Calibration in Multiple Language Detection with Large Performance Variation

Raymond W. M. Ng*, Cheung-Chi Leung†, Tan Lee*, Bin Ma† and Haizhou Li†‡

*Department of Electronic Engineering
The Chinese University of Hong Kong

†Institute for Infocomm Research
Singapore

‡ Department of Computer Science and Statistics,
University of Eastern Finland, Finland

ICASSP 2011
May 24, 2011

Contents

Contents

Score fusion and calibration

Score fusion and calibration combine and/or adjust the numerical value of the scores from one or multiple detector systems for a lower detection cost.

Multi-dimensional score vector $\longrightarrow$ Decision to detection problem (scalar)

Questions concerned:

- How to adjust (and combine) the numerical value of scores.
- Whether or not some criteria are used to guide the adjustment.

Common approaches to fusion and calibration

- Combination backend [Jain et al. 2005]
- LDA+Gaussian backend [Shen et al. 2006]
- Logistic regression backend [Brümmer et al. 2007]

These methods are/could be approximated by affine transformations.

Performance variation

Performance variation:

- among different detector systems.
- among different language detectors.

Contents

Performance variation among detector systems

Language detection systems can have large performance variation. For LRE 2009,

EER of phonotactic system:	3.54% $\pm$ 2.7%
EER of prosodic system:	19.40% $\pm$ 6.0%

We want to ...

- Investigate the parameter settings in MLR.
- Demonstrate error reduction of C_{avg} (with global error threshold) by the prosodic system.

Score fusion with 2 detector systems

Suppose we have 2 language detector systems:

Phonotactic-based system (*ph*)

Prosodic-based system (*pr*)

The likelihood scores to target language n_t (hypothesis) in trial k are $p_{ph}(k|n_t)$ and $p_{pr}(k|n_t)$.

Combination of system scores,

$$\log \hat{p}(k|n_t) = \log p_{ph}(k|n_t) + \beta \log p_{pr}(k|n_t) + \gamma_{n_t}.$$

Institute for Infocomm Research

A*STAR

Multi-class logistic regression (MLR)

In MLR, consider $\log\hat{p}(k|n_t) = \log p_{ph}(k|n_t) + \beta \log p_{pr}(k|n_t) + \gamma_{n_t}$. Parameter β and γ are optimized, with maximum-a-posteriori criterion,

$$\max_{\beta,\gamma} \sum_{n_t} \frac{1}{\|\mathcal{I}(n_t)\|} \sum_{k\in\mathcal{I}(n_t)} \log \frac{\exp\hat{p}(k|n_t)}{\sum_n \exp\hat{p}(k|n)}.$$

To cope with large performance variation,

- Language-specific β_{n_t} parameters will be used.
- MLR with and without the bias removing vector γ will be compared.

MLR parameter optimization is carried out by the multi-class FoCal toolkit, with a little code modification [Brümmer and du Preez 2006].

Contents

Performance variation among detectors

In LRE 2009, there are some pairs of related languages.
Detection to these related languages becomes a bottleneck.

- Russian-Ukrainian
- Hindi-Urdu
- Farsi-Dari
- Bosnian-Croatian
- English(American)-English(Indian)

While the average error is about 4% ...

For Bosnian: Error $=$ 20%
Confusion between Bosnian and Croatian $=$24%

For Hindi: Error $=$ 8%
Confusion between Hindi and Urdu $=$60%

Minimum erroneous deviation - Score Transformation

A calibration algorithm based on minimum erroneous deviation was proposed earlier [Ng et al. 2010].

Hypothesis: There are pairs of detectors which contain similar and complementary information.

- $\lambda^{n_t}_{\neg n_t}$,$\lambda^{n_r}_{\neg n_r}$: Log likelihood ratio of *target* and *related languages*.
- On top of MLR, we find optimal α_{n_t,n_r} where,

$$\lambda'^{n_t}_{\neg n_t}(k) = \lambda^{n_t}_{\neg n_t}(k) + \alpha_{n_t,n_r}\lambda^{n_r}_{\neg n_r}(k).$$

- Score transformation is affine, same as MLR.

Minimum erroneous deviation - Score Transformation

A calibration algorithm based on minimum erroneous deviation was proposed earlier [Ng et al. 2010].

Hypothesis: There are pairs of detectors which contain similar and complementary information.

- $\lambda^{n_t}_{\neg n_t}, \lambda^{n_r}_{\neg n_r}$: Log likelihood ratio of *target* and *related languages*.
- On top of MLR, we find optimal α_{n_t,n_r} where,

$$\lambda'^{n_t}_{\neg n_t}(k) = \lambda^{n_t}_{\neg n_t}(k) + \alpha_{n_t,n_r}\lambda^{n_r}_{\neg n_r}(k), k \in \{\tilde{\mathcal{I}}(n_t) \cup \tilde{\mathcal{I}}(n_r)\}.$$

- Score transformation is affine, same as MLR.
- MLR operates on global data set. The proposed calibration operates on selected data subset.

Minimum erroneous deviation - Parameter optimization

$$\min_{\upsilon,\alpha_{n_t,n_r}} \sum_{k\in\{\tilde{\mathcal{I}}(n_t)\cup\tilde{\mathcal{I}}(n_r)\}} \max\left[y_{n_t}(k)\times\left(\lambda'^{n_t}_{\neg n_t}(k)-(\theta+\upsilon)\right),0\right]$$

subject to (s.t.) $|\alpha_{n_t,n_r}| \leq 1$,

$$y_{n_t}(k)=\begin{cases}-(N-1) & \text{if } k\in\mathcal{I}(n_t).\\ 1 & \text{otherwise .}\end{cases}$$

- $\lambda'^{n_t}_{\neg n_t}(k)-(\theta+\upsilon)$: Deviation of $\lambda'^{n_t}_{\neg n_t}(k)$ from reference $\theta+\upsilon$.
- Product of $y_{n_t}(k)$ and the deviation:

  Positive for erroneous detection, negative for correct detection.
- Optimization minimizes total erroneous deviation.
- υ shifts the detection threshold. N scales the importance of misses and false alarms.

Institute for Infocomm Research

Comparison between MLR and Min erroneous deviation calibration

	Multi-class logistic regression (MLR)	Min erroneous deviation calibration
Same:	Affine transformation of score/llr	Affine transformation of score/llr

Comparison between MLR and Min erroneous deviation calibration

	Multi-class logistic regression (MLR)	Min erroneous deviation calibration
Same:	Affine transformation of score/llr	Affine transformation of score/llr
Different:	MAP criterion	Minimum erroneous deviation criterion

Comparison between MLR and Min erroneous deviation calibration

	Multi-class logistic regression (MLR)	Min erroneous deviation calibration
Same:	Affine transformation of score/llr	Affine transformation of score/llr
Different:	MAP criterion	Minimum erroneous deviation criterion
	Global data set operation	Selected data subset operation

Comparison between MLR and Min erroneous deviation calibration

	Multi-class logistic regression (MLR)	Min erroneous deviation calibration
Same:	Affine transformation of score/llr	Affine transformation of score/llr
Different:	MAP criterion	Minimum erroneous deviation criterion
	Global data set operation	Selected data subset operation
	Stand-alone process	Operated on top of MLR

Institute for Infocomm Research

A*STAR

Comparison between MLR and Min erroneous deviation calibration

	Multi-class logistic regression (MLR)	Min erroneous deviation calibration
Same:	Affine transformation of score/llr	Affine transformation of score/llr
Different:	MAP criterion	Minimum erroneous deviation criterion
	Global data set operation	Selected data subset operation
	Stand-alone process	Operated on top of MLR
	Application independent	Specific setting for υ, N

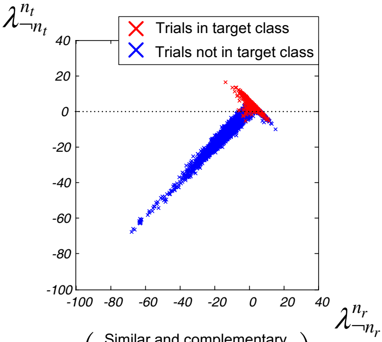
Shortcomings of the previous proposal

- In the earlier proposal, target languages to be calibrated has to be predetermined.
- We want to enhance the calibration algorithm by allowing on-the-fly selection of target languages for calibration.

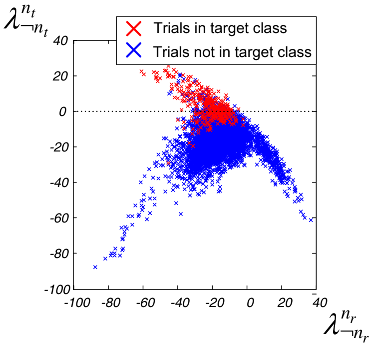
Analysis to pairs of language detectors

Hypothesis: Log likelihood ratios for n_t and n_r contain similar and complementary information.

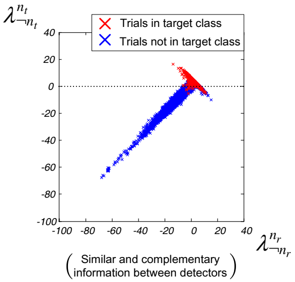
- Analyzing the $C_2^{23} = 253$ pairs of detectors...

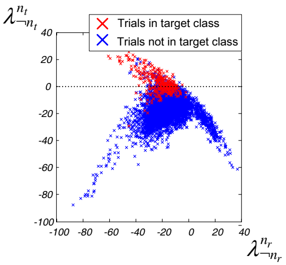


(Similar and complementary information between detectors)



Heuristics to choose pairs of detectors for calibration





Two heuristics are derived

- Minimum correlation of 0.9 to invoke the calibration mechanism.
- For every n_t, find the language with highest correlation as n_r.

Contents

Experimental setup

NIST LRE 2009 30-second close-set language detection

- Number of target languages: 23, Number of test trials: 10558
- Systems: Phonotactic PPRVSM (*ph*) [$C_{avg} = 4.69\%$] + Prosodic (*pr*)
- LDA+Guassian backend for each system

Experimental tasks

- Try different MLR parameters
- On-the-fly selection of n_t,n_r pairs for calibration
- Minimum erroneous deviation calibration
- Analysis of calibration results

Development set for MLR fusion and minimum erroneous deviation calibration: 6041 trials from LRE2007 and self-extracted VOA broadcast materials.

Fusion results with different MLR parameters

C_{avg} with *ph* system only is 4.69%. For fusion with *pr* system with different MLR settings:

	γ absent	γ present
Universal β	4.42%	4.24%
Language-dependent β_{n_t}	4.38%	4.20%

- Only marginal error reduction by language-dependent β_{n_t}.
- 10.5% relative reduction of C_{avg} for MLR with γ present.

(For "language-dependent EER", four MLR settings give similar errors.)

n_t,n_r pairs by correlation method

n_t	n_r	n_t	n_r
Amharic	Pashto	Hindi	Urdu
Bosnian	Croatian	Korean	Mandarin
Cantonese	Vietnamese	Mandarin	Vietnamese
Creole-Haitian	French	Pashto	Dari
Croatian	Bosnian	Portuguese	American English
Dari	Farsi	Russian	Spanish
American English	Indian English	Spanish	Indian English
Indian English	American English	Turkish	Pashto
Farsi	Dari	Ukrainian	Russian
French	Creole-Haitian	Urdu	Hindi
Georgian	Russian	Vietnamese	Cantonese
Hausa	French		

The correlation method recovers all language pairs which are specified as “mutually intelligible” languages in LRE 2009.
High correlation in the imposter data is a necessary but not sufficient condition for calibration algorithm to work effectively.

Minimum erroneous deviation calibration

With MED, C_{avg} reduces from 4.20% to 3.31%.

Looking into specific detectors,

$$C_{\text{avg}} = \frac{1}{N}\sum_{n_t=1}^{N} C_{\text{detect}}(n_t)$$

$$\text{where} \quad C_{\text{detect}}(n_t) = \frac{1}{2}P_{\text{Miss}}(n_t) + \sum_{n_n \neq n_t} \frac{1}{2}\frac{P_{\text{FA}}(n_t, n_n)}{N-1}$$

n_t	n_r	α_{n_t,n_r}	$P_{\text{Miss}}(n_t)$	$P_{\text{FA}}(n_t, n_r)$
Bosnian	Croatian	0.79	38.59% → 10.14%	24.20% → 78.19%
Hindi	Urdu	0.64	8.13% → 1.81%	59.89% → 95.78%
Ukrainian	Russian	0.71	22.16% → 11.08%	2.55% → 27.90%

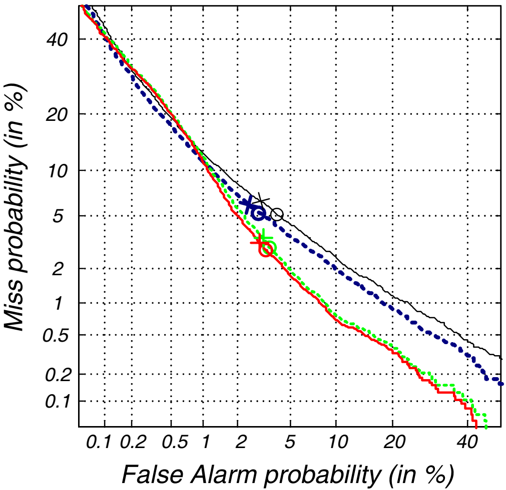
(Effective calibration with largest reduction of $C_{\text{detect}}(n_t)$)

n_t	n_r	α_{n_t,n_r}	$P_{\text{Miss}}(n_t)$	$P_{\text{FA}}(n_t, n_r)$
Cantonese	Vietnamese	−0.52	2.38% → 3.70%	6.03% → 2.86%
Creole-Haitian	French	0.61	1.24% → 0.93%	27.09% → 84.56%
French	Creole-Haitian	−0.67	3.04% → 6.58%	9.63% → 3.42%

(Non-effective calibration with largest increase of $C_{\text{detect}}(n_t)$)

Institute for Infocomm Research

Refining the calibration algorithm



System *ph* only
- ○ C_{min} = 4.58%
- × C_{avg} = 4.69%

Multi-class logistic regression fusion with system *pr*
- ○ C_{min} = 4.10%
- × C_{avg} = 4.20%

Calibration with minimum erroneous deviations (automatic n_t, n_r selection)
- ○ C_{min} = 3.24%
- × C_{avg} = 3.31%

Calibration with minimum erroneous deviations (automatic n_t, n_r selection & enforce $\alpha_{n_t,n_r} > 0$)
- ○ C_{min} = 3.06%
- × C_{avg} = 3.10%

Contents

Conclusion and Future Work

Parameter settings for multiple logistic regression with variation among detector systems

Enhancement of the minimum erroneous deviation calibration

- On-the-fly selection of related language pairs
- Extra optimization constraint in calibration algorithm to suppress detection misses

Future work: General applicability of the calibration algorithm

- Application on a normal data set without performance variation (LRE 2007)
- Calibration with multiple related languages
- More systematic methods in choosing the related languages

Reference

[Jain et al. 2005] A. Jain et al., “Score normalization in multimodal biometric systems,” *Pattern Recognition*, vol. 38, pp. 2270-2285, 2005.

[Shen et al. 2006] W. Shen et al., “Experiments with lattice-based PPRLM language identification,” in *Proc. Odyssey 2006*, 2006.

[Brümmer et al. 2007] N. Brümmer et al., “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *IEEE Trans. Audio, Speech, Lang. Pcs.*, vol. 15, no. 7, pp. 2072-2084, 2007.

[Ng et al. 2010] R.W.M. Ng et al., “Detection target dependent score calibration for language recognition,” in *Proc. Odyssey 2010*, pp. 91-96, 2010.

[Boyd 2004] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[Brümmer and du Preez 2007] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Comp., Speech and Lang.*, vol. 20, no. 2-3, pp. 230-275, 2006.

Appendix: Optimal MLR parameters

Language	β_{n_t}	γ	EER(*ph*)	EER(*pr*)	Language	β_{n_t}	γ	EER(*ph*)	EER(*pr*)
Amharic	0.61	-0.38	0.75%	15.08%	Hindi	0.59	-0.46	8.13%	23.79%
Bosnian	0.34	1.93	9.29%	25.07%	Korean	0.60	-0.54	1.30%	20.99%
Cantonese	0.52	-0.03	1.56%	10.06%	Mandarin	0.51	-0.25	1.16%	10.27%
Creole-Haitian	0.61	-0.42	2.12%	16.21%	Pashto	0.61	-0.51	4.62%	18.33%
Croatian	0.34	1.83	5.62%	23.95%	Portuguese	0.54	-0.55	1.26%	18.60%
Dari	0.49	0.23	8.73%	26.19%	Russian	0.63	-0.75	2.36%	22.99%
American English	0.54	-0.42	3.78%	24.53%	Spanish	0.53	-0.14	1.54%	25.51%
Indian English	0.50	-0.29	5.23%	13.11%	Turkish	0.59	0.21	1.27%	18.83%
Farsi	0.46	0.56	1.99%	23.33%	Ukrainian	0.63	0.10	6.67%	26.81%
French	0.63	-0.72	2.79%	17.78%	Urdu	0.60	-0.58	5.81%	25.85%
Georgian	0.61	0.11	1.54%	21.09%	Vietnamese	0.47	-0.11	2.54%	6.03%
Hausa	0.28	1.16	1.28%	11.86%					

Appendix: Optimization criteria and Data set involved

Language	MLR	Full data set MAP	Full data set MED	Selected data subset MAP	Selected data subset MED	Language	MLR	Full data set MAP	Full data set MED	Selected data subset MAP	Selected data subset MED
Amharic	0.63%	1.25%	0.70%	3.97%	0.76%	Hindi	7.53%	5.22%	4.84%	6.47%	5.05%
Bosnian	20.03%	6.76%	7.04%	8.52%	7.06%	Korean	0.82%	0.97%	0.82%	26.34%	0.82%
Cantonese	1.43%	1.43%	1.43%	1.43%	1.43%	Mandarin	0.97%	0.96%	0.97%	6.39%	0.97%
Creole-Haitian	1.60%	2.61%	2.61%	2.11%	2.69%	Pashto	3.53%	4.06%	3.02%	11.27%	3.23%
Croatian	9.31%	6.67%	8.90%	5.99%	6.48%	Portuguese	1.20%	1.26%	1.20%	6.46%	1.20%
Dari	8.48%	8.38%	8.48%	6.05%	6.01%	Russian	2.71%	2.55%	2.71%	6.07%	2.71%
American English	3.78%	3.78%	3.78%	3.78%	3.43%	Spanish	2.04%	2.19%	2.17%	2.91%	2.22%
Indian English	4.32%	3.04%	3.85%	3.93%	3.83%	Turkish	2.87%	2.87%	2.99%	3.11%	2.99%
Farsi	2.49%	2.49%	2.68%	2.49%	2.60%	Ukrainian	11.30%	8.91%	6.34%	9.05%	6.35%
French	2.32%	2.32%	2.32%	2.02%	2.32%	Urdu	5.20%	4.79%	5.14%	5.22%	5.03%
Georgian	1.31%	1.31%	1.30%	1.31%	1.34%	Vietnamese	2.02%	1.94%	2.02%	10.53%	2.03%
Hausa	0.66%	1.37%	0.66%	3.93%	0.66%						
						[All]	4.20%	3.35%	3.30%	6.06%	3.10%