Toward Text-independent Cross-lingual Speaker Recognition Using English-Mandarin-Taiwanese Dataset

Yi-Chieh Wu and Wen-Hung Liao

Dept. of Computer Science, National Chengchi University, Taipei, TAIWAN

Introduction

Over 40% of the world's population is bilingual. Existing speaker identification/verification systems, however, assume the same language type for both enrollment and recognition stages. Consider the possible factors affecting the performance of a text-independent speaker recognition (TISR) system, we speculate if the language employed plays an important role. In this work, we investigate the feasibility of employing multilingual speech for biometric applications. We establish a dataset containing audio recorded in English, Mandarin and Taiwanese (6 females and 10 males, all native Mandarin speakers), as shown in Table 2.

The features, namely, i-vector[1], d-vector[2] and x-vector[3] have been evaluated for both speaker verification (SV) and identification (SI) tasks. The SI result of features comparison is shown in Table 1. Preliminary experimental results indicate that x-vector achieves the best overall performance.

Table 1:Pilot Study: Feature comparison of speaker identification using SVM

Model	English							
Feature	Eng.		Man.		Twn.		RQ.	
	train (N)	*test*	*train*	*test*	*train*	*test*	*train*	*test*
d-vec.(N16/M10)	91.94% (19634)	90.32%	78.25%	73.05%	68.19%	71.3%	70.01%	74.75%
d-vec.(N64/M10)	90.42% (19580)	88.73%	72.98%	66.17%	64.15%	65.05%	70.06%	72.94%
i-vec.(3s)	100% (9467)	99.62%	92.22%	91.51%	74.71%	75%	87.02%	88.73%
x-vec.(Origin)	100% (1773)	**100%**	**100%**	**100%**	88.96%	89.27%	**100%**	**100%**
x-vec.(3s)	100% (9467)	99.96%	98.61%	98.77%	**92.69%**	**92.41%**	96.74%	99.82%
Model	Mandarin							
Feature	Eng.		Man.		Twn.		RQ.	
	train	*test*	*train (N)*	*test*	*train*	*test*	*train*	*test*
d-vec.(N16/M10)	65%	64.82%	95.32% (6004)	91.12%	70.8%	75.19%	65.74%	70.38%
d-vec.(N64/M10)	58.86%	57.06%	92.57% (6261)	88.98%	66.19%	66.36%	62.62%	64.91%
i-vec.(3s)	86.3%	87.77%	100% (5106)	99.56%	76.97%	79.07%	86.38%	86.55%
x-vec.(Origin)	93.91%	94.05%	100% (233)	**100%**	95.02%	94.92%	**100%**	**100%**
x-vec.(3s)	**96.62%**	**96.45%**	100% (5106)	99.91%	**96.66%**	**96.67%**	98.08%	96.55%
Model	Taiwanese							
Feature	Eng.		Man.		Twn.		RQ.	
	train	*test*	*train*	*test*	*train (N)*	*test*	*train*	*test*
d-vec.(N16/M10)	64.04%	64.8%	79.23%	74.08%	91.86% (6451)	87.04%	64.25%	71.17%
d-vec.(N64/M10)	58.81%	58.5%	75.37%	72.17%	88.37% (6466)	85.17%	63.54%	64.22%
i-vec.(3s)	81.93%	81.35%	86.92%	90.19%	100% (2653)	96.67%	80.82%	81.45%
x-vec.(Origin)	**98.14%**	**97.71%**	**100%**	98.08%	100% (743)	**100%**	**100%**	**100%**
x-vec.(3s)	95.33%	96.11%	98.79%	**99.82%**	100% (2653)	**100%**	98.08%	98.91%
Model	Random Questions							
Feature	Eng.		Man.		Twn.		RQ.	
	train	*test*	*train*	*test*	*train*	*test*	*train (N)*	*test*
d-vec.(N16/M10)	57.32%	58.79%	66.39%	63.99%	60.16%	64.51%	91.11% (2204)	87.08%
d-vec.(N64/M10)	51.19%	51.44%	59.88%	55.29%	53.42%	52.77%	86.22% (2271)	80.05%
i-vec.(3s)	82.94%	83.96%	87.27%	88%	70.98%	73.52%	100% (1872)	93.45%
x-vec.(Origin)	**95.6%**	**95.65%**	**98.71%**	**98.08%**	90.17%	91.53%	100% (64)	**100%**
x-vec.(3s)	94.57%	95.04%	98.55%	98.07%	**91.71%**	**92.96%**	100% (1872)	**100%**

Methodology

Fig.1 depicts a typical training/test process of the TISR architecture. To evaluate cross-lingual performance, we adopt SVM classifiers[4] in SI tasks, and PLDA classifiers[5] in SV tasks. Moreover, one speaker is selected as leave-one-out (un-enrolled). At last, a cross-lingual model trained by all language data is built to find out whether hybrid training is beneficial.

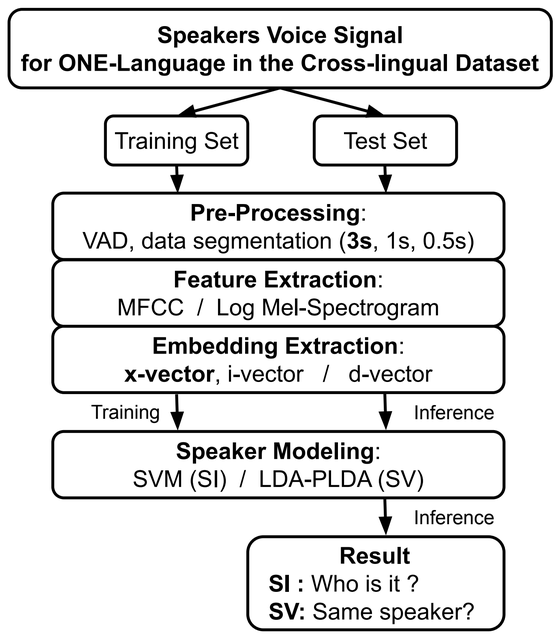

Figure 1:Workflow for each language round

Contact Information

● Wen-Hung Liao: whliao@nccu.edu.tw

English-Mandarin-Taiwanese Dataset

Table 2:Overview of English-Mandarin-Taiwanese Dataset

Language	Number of Utterances	Total Length (minute)	Length Per Utt. μ (σ) (second)
English	2210	$\approx$ 202	5.79 (1.89)
Mandarin	285	$\approx$ 60	14.1 (5.16)
Taiwanese	920	$\approx$ 67	5.39 (3.43)
Random Questions	80	$\approx$ 22	16.75 (4.68)

Experimental Results

The cross-lingual results of SVM speaker models using original and 3-second audio are shown in Table 3 and 4. The worst results are marked in red. More results are available at: `http://www.cs.nccu.edu.tw/~d10402/icpr2020.html`.

Table 3:Metric results of SVM models using original data

Model	Eng. Train			Eng. Test			Man. Train			Man. Test		
	Acc.	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*
Eng.	100%	0.05	1	100%	0.065	1	100%	0.204	1	100%	0.224	1
Man.	93.54%	1.158	0.941	93.69%	1.157	0.942	100%	0.447	1	100%	0.518	1
Twn.	98.14%	0.441	0.984	97.82%	0.437	0.98	100%	0.212	1	97.87%	0.229	0.983
RQ.	95.33%	1.643	0.952	95.63%	1.636	0.953	98.58%	1.416	0.98	97.87%	1.468	0.973
Mix.	100%	0.027	1	100%	0.034	1	100%	0.024	1	100%	0.037	1
Model	**Twn. Train**			**Twn. Test**			**RQ. Train**			**RQ. Test**		
	Acc.	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*
Eng.	88.52%	0.52	0.899	89.16%	0.54	0.897	100%	0.181	1	100%	0.146	1
Man.	95.55%	1.167	0.955	95.18%	1.167	0.955	100%	0.859	1	100%	0.867	1
Twn.	100%	0.106	1	100%	0.166	1	100%	0.331	1	100%	0.319	1
RQ.	89.96%	1.737	0.888	91.57%	1.737	0.913	100%	1.117	1	100%	1.216	1
Mix.	100%	0.038	1	98.8%	0.085	0.99	100%	0.056	1	100%	0.04	1

Table 4:Metric results of SVM models using 3-second audio

Model	Eng. Train			Eng. Test			Man. Train			Man. Test		
	Acc.	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*
Eng.	100%	0.009	1	99.96%	0.017	1	98.53%	0.147	0.984	98.7%	0.148	0.987
Man.	96.57%	0.295	0.956	96.43%	0.305	0.949	100%	0.019	1	99.91%	0.039	0.999
Twn.	95.27%	0.356	0.957	96.04%	0.354	0.96	98.72%	0.167	0.983	99.81%	0.137	0.998
RQ.	94.53%	0.494	0.953	94.95%	0.502	0.953	98.47%	0.344	0.983	98.14%	0.341	0.978
Mix.	100%	0.005	1	100%	0.009	1	100%	0.005	1	100%	0.01	1
Model	**Twn. Train**			**Twn. Test**			**RQ. Train**			**RQ. Test**		
	Acc.	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*	*Acc.*	*L-L*	*a-F1*
Eng.	92.84%	0.302	0.89	92.25%	0.313	0.88	96.68%	0.167	0.957	99.81%	0.117	0.996
Man.	96.8%	0.248	0.946	96.6%	0.257	0.949	98%	0.245	0.976	96.3%	0.266	0.964
Twn.	100%	0.032	1	100%	0.065	1	98%	0.261	0.978	98.83%	0.219	0.987
RQ.	91.91%	0.523	0.904	93.19%	0.503	0.918	100%	0.054	1	100%	0.148	1
Mix.	100%	0.007	1	100%	0.023	1	98.62%	0.063	0.984	100%	0.029	1

Conclusion

We investigated the usability of cross-lingual speech on TISR tasks through a multilingual dataset containing English, Mandarin and Taiwanese speech. We conducted a pilot study to evaluate the state-of-the-art acoustic features. The result showed that x-vector is a potential candidate for cross-lingual representation.

In SI tasks, we achieved over 91% cross-lingual accuracy on all models using 3-second audio. In SV tasks, the EER among cross-lingual test is at most 6.52% on the model trained using English corpus. For the analysis of leave-one-out and enrolled users, the results indicate that there exists a certain degree of individual differences. The same speaker may perform very differently between reading aloud and answering random questions, even Mandarin is spoken in both scenarios.

References

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[2] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[3] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Sergey Ioffe. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer, 2006.