SNR-Invariant Multi-Task Deep Neural Networks for Robust Speaker Verification

SPL Journal Presentation

Qi YAO and Man-Wai MAK

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University

Introduction

- We observed that background noise in utterances will not only enlarge the speaker-dependent i-vector clusters but also shift the clusters, with the amount of shift depending on the signal-to-noise ratio (SNR) of the utterances;
- We propose to utilize clean i-vectors as well as available meta information to train a hierarchical regression DNN (H-RDNN) and a multitask DNN (MT-DNN);
- We show that the proposed DNN architecture together with the PLDA backend outperform the multi-condition PLDA model and mixtures of PLDA in noisy environments.

Proposed Models

- ***Hierarchical regression DNN:***
  - The first regression DNN is trained to map noisy i-vectors to their respective speaker-dependent cluster means of clean i-vectors:

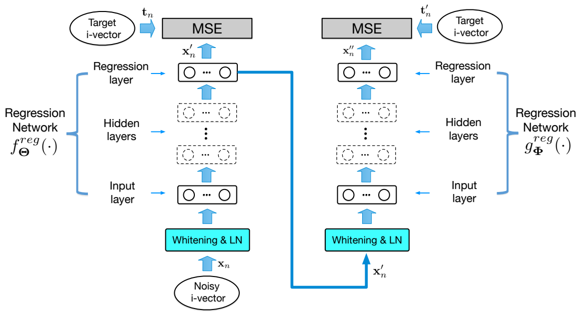
Stage 1:
$$\min_{\Theta} \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\|f_{\Theta}^{reg}(\mathbf{x}_n)-\mathbf{t}_n\|_2^2+\frac{\beta_{reg_1}}{2}\|\Theta\|_2^2$$

where $\mathbf{x}_n$ is the n-th training i-vector pre-processed by WCCN and LN; $\mathbf{t}_n$ is the corresponding target i-vector obtained by averaging speaker-dependent i-vectors from clean utterances;

  - The second regression DNN is trained to regularize the outliers that cannot be denoised properly by the first regression DNN:

Stage 2:
$$\min_{\Phi} \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\|g_{\Phi}^{reg}(\mathbf{x}'_n)-\mathbf{t}'_n\|_2^2+\frac{\beta_{reg_2}}{2}\|\Phi\|_2^2$$

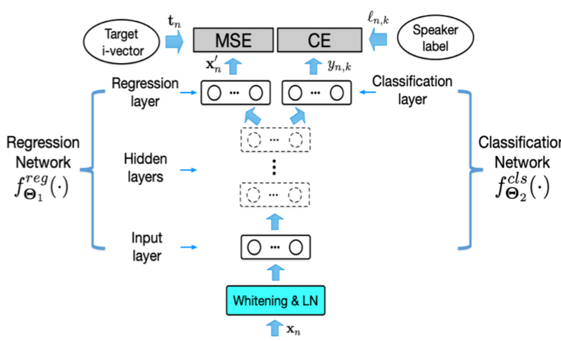
where $\mathbf{x}'_n$ is the n-th i-vector denoised by the first DNN; $\mathbf{t}'_n$ is the corresponding i-vector from the original i-vector set (no noise corruption) and then denoised by the first DNN.



- ***Multi-task DNN:***
  - To reduce the speaker information loss in the regression task, we introduce a second task-specific layer at the top of the regression network (1-st stage) to classify speakers:

$$\min_{\Theta_2} -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\ell_{n,k}\log y_{n,k}+\frac{\beta_{cls}}{2}\|\Theta_2\|_2^2$$

where $\ell_{n,k}$ is the k-th element of ℓ_n; if the utterance of $\mathbf{x}_n$ is spoken by the k-th speaker, then $\ell_{n,k}=1$, otherwise it is equal to 0; $y_{n,k}$ is the posterior probability of the k-th speaker.



Results

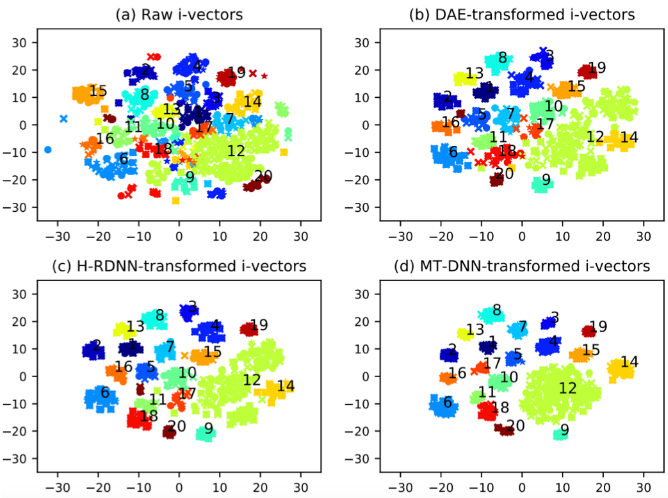
Performance on the original test segments in NIST 2012 SRE, with babble noise at SNR of 0dB, 6dB and 15dB being added to the training utterances.

Model	CC4		CC5	
	EER	minDCF	EER	minDCF
Multi-condition PLDA	4.02	0.352	3.61	0.343
SI-mPLDA	3.88	0.333	3.21	0.306
SD-mPLDA	3.80	0.353	3.48	0.338
DAE+PLDA	3.32	0.339	2.93	0.329
H-RDNN+PLDA	3.24	0.348	2.95	0.338
MT-DNN+PLDA	**3.12**	**3.325**	**2.76**	**0.307**

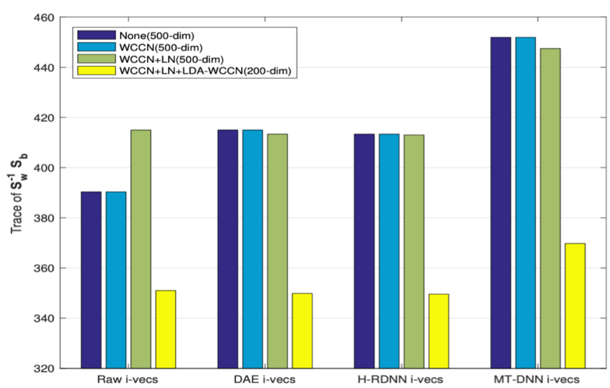
Performance in CC4 of NIST 2012 SRE under 3 SNR conditions in the test segments. The results below only show the case in which babble noise was added to 3 SNR test sets. For full results, refer to Reference 1.

Model	15 dB		6 dB		0 dB	
	EER	minDCF	EER	minDCF	EER	minDCF
Multi-condition PLDA	2.54	0.266	2.84	0.325	4.56	0.500
SI-mPLDA	2.42	**0.237**	2.85	**0.314**	4.55	0.478
SD-mPLDA	2.68	0.271	2.91	0.335	4.36	0.497
DAE+PLDA	2.13	0.278	2.55	0.337	3.89	0.437
H-RDNN+PLDA	2.15	0.280	2.56	0.341	3.92	0.435
MT-DNN+PLDA	**2.05**	0.272	**2.48**	0.316	**3.82**	**0.428**

T-SNE plots of 20 speaker clusters from 3 SNR groups (org+15dB+6dB, telephone speech, babble noise). The raw i-vectors in (a) were transformed by DAE (b), H-RDNN (c), and MT-DNN (d). Speakers are marked with different colors and i-vectors from the three SNR groups are marked with ∘, ×, and ∗, respectively.



Dispersion of 20 speaker clusters from 3 SNR groups (org+15dB+6dB, telephone speech, babble noise). The x-axis indicates the types of DNN transformation methods applied to the raw i-vectors. The y-axis indicates the values of $Tr(\mathbf{S}_w^{-1}\mathbf{S}_b)$. The colors in the legend denotes different i-vector post-processing methods applied to the DNN-transformed i-vectors.



Conclusions

- The compactness of speaker-dependent i-vector clusters largely depends on the SNR of utterance;
- Meta information, such as the speaker identity of utterance, helps MT-DNN to discriminate i-vectors from different speakers while perform the denoising task.

References

1. Q. Yao and M.W. Mak, "SNR-Invariant Multi-Task Deep Neural Networks for Robust Speaker Verification, ***IEEE Signal Processing Letters***, vol. 25, no. 11, pp. 1670-1674, Nov. 2018.
2. Na Li and M.W. Mak, "SNR-Invariant PLDA Modeling in Nonparametric Subspace for Robust Speaker Verification", ***IEEE/ACM Trans. on Audio Speech and Language Processing***, vol. 23, no. 10, pp. 1648-1659, Oct. 2015..