論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Speaker verication using deep speaker-discriminative representations
著者(和文)	Price RyanWilliam, 篠田 浩一
Authors(English)	Ryan Price, Koichi Shinoda
出典(和文)	日本音響学会講演論文集, , , pp. 81-82
Citation(English)	2013 Spring Meeting ASJ, , , pp. 81-82
発行日 / Pub. date	2013, 3

Speaker verification using deep speaker-discriminative representations

◎Ryan Price and Koichi Shinoda (Tokyo Institute of Technology) *

1 Introduction

Gaussian mixture model support vector machines (GMM-SVM) with MFCC input features are widely used for speaker verification (SV)[3]. However, the generative approach of GMM speaker modeling lacks the ability to extract speaker specific information by discriminative means and also requires learning a speaker independent universal background model (UBM) that covers the variability in the training data significantly well. Recent studies (eg. [1]) have shown that regularized siamese deep networks (RSDN) have promise as an effective method for extracting speaker specific information from a spectral representation but they have not yet been successfully paired with a robust approach to speaker modeling and decision-making for SV. We address this by combining a GMM-SVM system with speaker specific input features extracted from a discriminatively trained RSDN.

2 Regularized Siamese Deep Network

Unlike standard deep architectures trained with speech input, RSDN learns a speaker specific representation by discriminatively training a subset of the hidden units using pairs of speech segments (Fig. 1). The RSDN is built upon the concept of denoising autoencoders [2] and training consists of unsupervised pretraining followed by supervised finetuning. During the pretraining phase, unlabeled speech data, such as MFCC features, are used for training the network. Raw input is corrupted by Gaussian noise and the network is trained to reconstruct the clean input similar to [2]. During finetuning, pairs of short speech segments (X_1, X_2), coming either from the same speaker (genuine pairs) or from different speakers (imposter pairs), are presented to the network. The contrastive loss function (eq.(1)) is applied to a subset of the code layer units in order to learn a speaker specific representation. The network is also trained to minimize reconstruction loss (eq.(2)), ensuring the remaining neurons in the

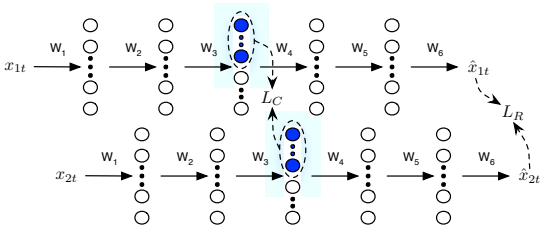

Fig. 1 Regularized Siamese Deep Network

network act as a regularizer. The contrastive loss and reconstruction loss are combined in the overall loss function (eq.(3)).

$$C_m = \|\boldsymbol{\mu}_{S1} - \boldsymbol{\mu}_{S2}\|_2^2 \qquad C_s = \|\Sigma_{S1} - \Sigma_{S2}\|_F^2$$

$$L_C(X_1, X_2) = \mathcal{I}\big[C_m + C_s\big] + (1-\mathcal{I})\big[e^{-\frac{C_m}{\lambda_\mu}} + e^{-\frac{C_s}{\lambda_{\text{cov}}}}\big] \quad (1)$$

$$L_R(X_1, X_2) = \frac{1}{T}\sum_{t=1}^{T}\big[\|x_{1t} - \hat{x}_{1t}\|_2^2 + \|x_{2t} - \hat{x}_{2t}\|_2^2\big] \quad (2)$$

$$L(X_1, X_2) = \alpha L_R + (1-\alpha) L_C \quad (3)$$

where α determines the tradeoff between L_R and L_C, $\mathcal{I} = 1$ for genuine pairs and 0 for imposter pairs, λ_μ and λ_{cov} are tolerance bounds estimated from the training data, and $\boldsymbol{\mu}_{S1}$, $\boldsymbol{\mu}_{S2}$, Σ_{S1} and Σ_{S2} are the means and covariance matrices of the outputs of speaker specific code layer units corresponding to the segment pair (X_1, X_2), respectively.

3 Hybrid RDSN GMM-SVM

In previous work [1], RSDN code layer outputs were used to train single Gaussian speaker models from short speech segments. Scores for binary classification were calculated using the symmetric Gaussian log likelihood measure for a given test trial. While that approach offered promising results in the SV tasks studied, we believe that the speaker specific features extracted from the RSDN code layer can be combined with a more robust speaker modeling approach and classifier. We propose a straightforward approach to using RSDN representations along with GMM-SVM to form a hybrid system for SV.

After pretraining and finetuning the RSDN as described in Section 2, the RSDN parameters are

---

*話者同定のための識別的 Deep モデリング
プライス ライアン, 篠田 浩一 (東工大)

fixed and training the hybrid RSDN GMM-SVM system closely follows the typical GMM-SVM training procedure[3] except that the RSDN is used to extract speaker specific features from MFCC features for each input utterance. RSDN code layer outputs are extracted for each frame in the training data and are used to derive a GMM UBM. RSDN code layer outputs are also extracted for all frames in the enrollment and test data and then GMM supervectors are created on a per utterance basis using MAP adaptation of the means with a relevance factor of 1. GMM supervectors extracted from the utterances used to train the UBM are used as imposter examples to train an SVM model with a linear kernel for each target speaker in the enrollment set. Finally, scores are calculated for each target and nontarget trial using the target speaker's SVM model and the supervector extracted from the test utterance.

4 Experiment

We consider the problem of text-independent SV for evaluating the hybrid RSDN GMM-SVM system. The 242 male speakers from the NIST SRE 2004 1-side [4] training files were selected for training the RSDN and the GMM UBM. Utterances from 50 randomly selected male speakers who did not appear in the training data were taken from NIST SRE 2004 8-side [4] training files and used for development. For evaluation we randomly selected 100 male speakers from the NIST SRE 2006 1conv4w-1conv4w [4] task and use all trials associated with those 100 target speakers. In total, 453 genuine trials and 6057 imposter trials were used for evaluation. Silence was removed using an energy based VAD and a 19-dimensional MFCC vector is extracted every 10 ms using a 25 ms Hamming window.

The RSDN was implemented on a GPU and details of training are as follows. All frames in the training data were used for pretraining. For finetuning, we created approximately 3000 genuine pairs and 3000 imposter pairs of segments that are 500 frames in length[1]. We used a network with 3 hidden layers having sizes of 100, 100, and 200 hidden units, respectively. We used 100 speaker specific units in the code layer, which is half the number of hidden units in that layer, as in [1]. Learning rates of 0.01 and 0.001 were used for pretraining and finetuning, respectively. Other parameters were $\alpha = 0.2$, $\lambda_\mu = 100$ and $\lambda_{\text{cov}} = 2.5$. Early stopping was used with the development data to prevent overfitting.

4.1 Results

We use equal error rate (EER) as an evaluation metric. The number of mixture components was varied from 32 to 512 and best results for both systems are shown in table 1. The hybrid RSDN GMM-SVM system achieves about 5% relative improvement over the baseline GMM-SVM system.

Table 1 *EER(%) for MFCC based GMM-SVM and Hybrid RSDN GMM-SVM.*

System	Mix Comp	EER
GMM-SVM (MFCC input)	128	13.24
Hybrid RSDN GMM-SVM	64	**12.58**

5 Conclusions

We have demonstrated the novel combination of RSDN and GMM-SVM for text-independent SV. This hybrid system outperformed the baseline GMM-SVM system using MFCC input features for the subset of the NIST SRE 2006 task studied. In the future, we would like to apply it to larger SV tasks and combine it with channel compensation techniques.

References

[1] K. Chen and A. Salman, In *Advances in Neural Information Processing Systems* 24, 2011.

[2] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, In *Proc. ICML*, pp. 1096-1103, 2008.

[3] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, *IEEE Signal Processing Letters*, vol. 13, no. 5, May 2006.

[4] NIST Speaker Recognition Evaluation. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/.

[1] Mini-batch sizes were 100 frames and 500 frames for pretraining and finetuning, respectively. Note that eq.(1) is defined using 1st and 2nd order statistics of a speech segment that is T frames in length and it is thus necessary to use a mini-batch size of T for finetuning.