Regularized Auto-Associative Neural Networks for Speaker Verification

Sri Garimella, *Student Member, IEEE*, Sri Harish Mallidi, and Hynek Hermansky, *Fellow, IEEE*

***Abstract*—Auto-Associative Neural Network (AANN) is a fully connected feed-forward neural network, trained to reconstruct its input at its output through a hidden compression layer. AANNs are used to model speakers in speaker verification, where a speaker-specific AANN model is obtained by adapting (or retraining) the Universal Background Model (UBM) AANN, an AANN trained on multiple held out speakers, using corresponding speaker data. When the amount of speaker data is limited, this adaptation procedure leads to overfitting. Additionally, the resultant speaker-specific parameters become noisy due to outliers in data. Thus, we propose to regularize the parameters of an AANN during speaker adaptation. A closed-form expression for updating the parameters is derived. Further, these speaker-specific AANN parameters are directly used as features in linear discriminant analysis (LDA)/ probabilistic discriminant (PLDA) analysis based speaker verification system. The proposed speaker verification system outperforms the previously proposed weighted least squares (WLS) based AANN speaker verification system on NIST-08 speaker recognition evaluation (SRE). Moreover, the proposed speaker verification system obviates the need for an intermediate dimensionality reduction (or i-vector extraction) step.**

***Index Terms*—Adaptation, auto-associative neural network, regularization, speaker verification.**

I. Introduction

THE goal of a speaker verification is to verify whether a given utterance belongs to a claimed speaker or not based on a sample utterance from claimed speaker. In other words, the task is to verify whether a given two utterances of a speaker verification trial belong to the same speaker or not. Traditional speaker verification systems use likelihood ratio between Gaussian Mixture Model (GMM) based Universal Background Model (UBM) and its maximum *a posteriori* (MAP) adapted speaker-specific model for making decision [1].

AANN is used as an alternative to GMM for modeling the distribution of data [2], and it has several advantages—it relaxes the assumption of feature vectors to be locally normal and can capture higher order moments. An AANN is a fully connected feed-forward neural network with a hidden compression layer, and trained for auto-encoding (reconstructing its input at its output) task [3]. A block schematic of an AANN is shown in the Fig. 1. This architecture consists of three non-linear hidden layers between the linear input and output layers. The second hidden layer contains fewer nodes than the input layer, and is known as the compression layer.

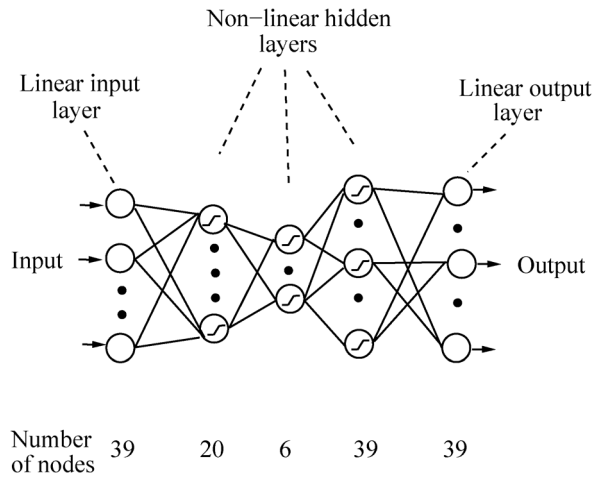

Fig. 1. Block schematic of an AANN.

Earlier AANN based speaker verification systems [2], [4], [5] use the reconstruction error difference computed using the UBM-AANN and the speaker-specific AANN models as a score for making decision. The UBM-AANN is obtained by training an AANN on multiple held out speakers using the stochastic gradient descent, where gradient is computed using the error back-propagation algorithm. Where as the speaker-specific AANN is obtained by adapting (or retraining) the entire UBM-AANN using corresponding speaker data. Better results are observed when only the weights connecting third hidden layer and output layer of an UBM-AANN are adapted. This indicates that adapting the entire UBM-AANN with limited amount of speaker data leads to overfitting. Additionally, the speaker-specific AANN parameters become noisy due to outliers in the data. These issues are addressed in [6] by projecting the adapted speaker-specific weights onto a low-dimensional subspace ($\mathbf{T}$ matrix), which is learned to minimize the reconstruction error between the speaker-specific weights and their projection in a WLS sense. The subspace coordinates representing the projection is known as an i-vector. The block diagram of WLS based AANN speaker verification system is shown in Fig. 2.

In this paper, we propose to regularize the weights connecting third hidden layer and output layer of an UBM-AANN during speaker adaptation. A closed-form expression for updating the weights is also derived. This obviates the need for further

Manuscript received July 10, 2012; revised September 01, 2012; accepted September 12, 2012. Date of publication October 02, 2012; date of current version October 15, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ali Taylan Cemgil.

The authors are with the ECE Department and the Center for Language and Speech Processing, Johns Hopkins University, Baltimore MD 21218 USA (e-mail: sivaram@jhu.edu; mallidi@jhu.edu; hynek@jhu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2012.2221706

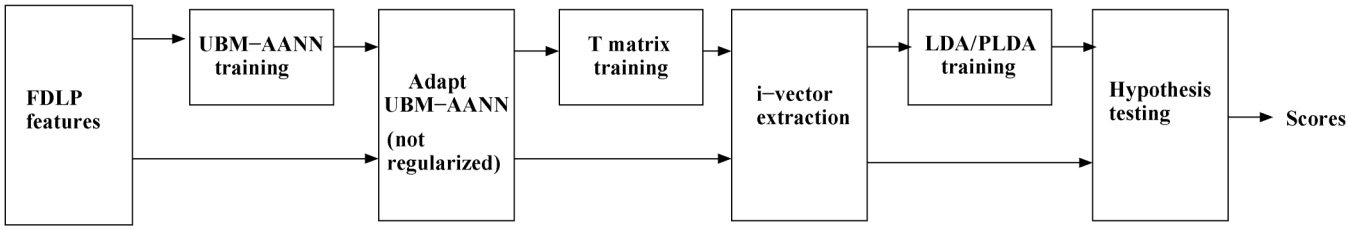

Fig. 2. Baseline WLS based AANN speaker verification system.

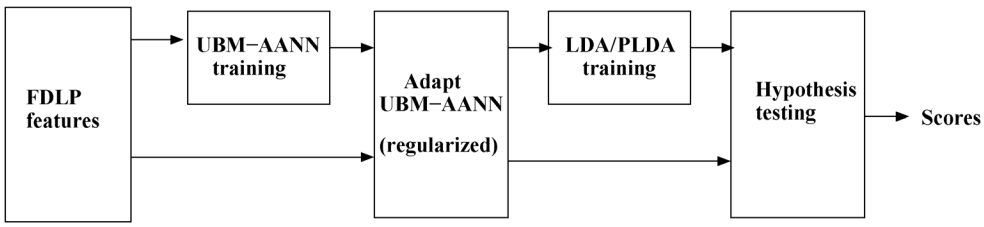

Fig. 3. Proposed regularized AANN based speaker verification system.

projecting these weights onto a low-dimensional subspace. We propose an AANN based speaker verification system where regularized speaker-specific weights are directly used as features for the linear discriminant analysis (LDA) followed by PLDA model, as shown in Fig. 3. Experimental results on NIST-08 SRE show that the proposed system outperforms the WLS based AANN speaker verification system[1] [6].

II. Regularized Adaptation of UBM-AANN

The weight matrix connecting third hidden layer and output layer of UBM-AANN is adapted for each utterance to obtain a speaker-specific model. Let $\mathbf{W}_{l,s}$ denote the adapted speaker-specific weight matrix corresponding to l^{th} session of s^{th} speaker. The output bias vector $\mathbf{b}$ of UBM-AANN is not adapted.

Let $\mathbf{f}_{i,l,s}$ be the i^{th} feature vector (frame) of an utterance corresponding to l^{th} session of s^{th} speaker, and $n(l,s)$ be the number of such frames in that utterance. The third hidden layer output vector of UBM-AANN corresponding to this input is denoted with $\mathbf{h}_{i,l,s}$. The following loss function (1) is minimized to obtain the speaker-specific weight matrix $\mathbf{W}_{l,s}$. It consists of two terms. The first term is the sum of squared reconstruction errors of the speaker-specific AANN. Second term represents the L^2 regularization of speaker-specific weights, where β is non-negative and controls the amount of regularization.

$$L(\mathbf{W}_{l,s}) = \sum_{i=1}^{n(l,s)} \|\mathbf{f}_{i,l,s} - \mathbf{b} - \mathbf{W}_{l,s}\mathbf{h}_{i,l,s}\|_2^2 + \beta n(l,s)\mathrm{tr}\left(\mathbf{W}_{l,s}\mathbf{W}_{l,s}^T\right). \quad (1)$$

[1]We found out that empirically there is no advantage of using mixture of AANNs over single AANN when speaker-specific weights are projected onto a subspace.

It is possible to derive the closed-form expression for weight matrix $\mathbf{W}_{l,s}$ by differentiating the expression above with respect to $\mathbf{W}_{l,s}$ and setting it to zero.

$$\frac{\partial L(\mathbf{W}_{l,s})}{\partial \mathbf{W}_{l,s}} = \mathbf{0} \Rightarrow$$

$$\sum_{i=1}^{n(l,s)} \left[\left(\mathbf{h}_{i,l,s}\mathbf{h}_{i,l,s}^T + \beta\mathbf{I}\right)\mathbf{W}_{l,s}^T - \mathbf{h}_{i,l,s}(\mathbf{f}_{i,l,s} - \mathbf{b})^T\right] = \mathbf{0} \Rightarrow$$

$$\mathbf{W}_{l,s} = \left[\sum_{i=1}^{n(l,s)} (\mathbf{f}_{i,l,s} - \mathbf{b})\mathbf{h}_{i,l,s}^T\right]\left[\sum_{i=1}^{n(l,s)} \left(\mathbf{h}_{i,l,s}\mathbf{h}_{i,l,s}^T + \beta\mathbf{I}\right)\right]^{-1}. \quad (2)$$

III. Regularized AANN Based Speaker Verification System

The block diagram of the proposed regularized AANN based speaker verification system is shown in Fig. 3. The various components of this system are described below.

A. Feature Extraction

The acoustic features used in our experiments are 39 dimensional frequency domain linear prediction (FDLP) features [7]–[11]. In this technique, sub-band temporal envelopes of speech are first estimated in narrow sub-bands (96 linear bands). These sub-band envelopes are then gain normalized to remove reverberation and channel artifacts. After normalization, the frequency axis is warped to 37 Mel bands in the frequency range of 125–3800 Hz to derive a gain normalized mel scale energy representation of speech. This is similar to the mel spectrogram obtained in conventional mel frequency cepstral coefficients (MFCC) feature extraction. These mel band energies are converted to cepstral coefficients by applying a log and Discrete Cosine Transform (DCT). The top 13 cepstral coefficients along with derivative and acceleration components are used as features, yielding 39 dimensional feature vectors. Finally, a subset of these feature vectors corresponding to speech are selected based on the voice activity detection information provided by NIST.

B. UBM-AANN

Gender-specific AANN based UBMs are trained on a telephone development data set consisting of audio from the NIST 2004 speaker recognition database, the Switchboard-2 Phase III corpora and the NIST 2005 speaker recognition database. We use only 400 male and 400 female utterances each corresponding to about 17 hours of speech.

AANN based UBMs are trained using the FDLP features (see Section III-A) to minimize the reconstruction error loss function [6]. Each UBM has a linear input and linear output layers along with three nonlinear (tanh nonlinearity) hidden layers. Both input and output layers have 39 nodes corresponding to the dimensionality of the input FDLP features. First, second and third hidden layers have 20, 6 and 39 nodes respectively. The number of nodes in each hidden layer is optimized for speaker verification task by fixing rest of the AANN configuration. Schematic of an AANN with this architecture is shown in the Fig. 1. We have modified the Quicknet package for training the AANNs [12].

C. Adaptation of UBM-AANN

The weight matrix ($39 \times 39 = 1521$ elements) connecting third hidden layer and output layer of a gender-specific UBM-AANN is adapted for each utterance to obtain a speaker specific weights. The closed-form expression for adapting these weights with regularization is described in Section II.

D. LDA/PLDA Training

Gender dependent linear discriminant analysis (LDA) transforms are trained to project vectorized adapted speaker-specific weight matrices onto a low-dimensional (240) space. The resultant low-dimensional projections are length normalized to reduce the mismatch during training and testing [13]. The development data for training consists of Switchboard-2, Phases II and III; Switchboard Cellular, Parts 1 and 2 and NIST 2004–2005 SRE [14]. The total number of male and female utterances is 12266 and 14936 respectively.

Subsequently, the length normalized vectors (denoted with $\mathbf{q}_{l,s}$) are modeled using the PLDA, a generative model for observations [15], [16]. They are assumed to be generated as

$$\mathbf{q}_{l,s} = \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\beta}_s + \boldsymbol{\epsilon}_{l,s}, \tag{3}$$

where $\boldsymbol{\mu}$ is an offset; $\boldsymbol{\Phi}$ is a matrix fewer columns than rows representing a low-dimensional subspace; $\boldsymbol{\beta}_s$ is a latent identity variable having a normal distribution with mean zero and covariance matrix identity; and $\boldsymbol{\epsilon}_{l,s}$ is a residual noise term assumed to be Gaussian with mean zero and full covariance matrix $\boldsymbol{\Sigma}_\epsilon$. Additionally, these variables are assumed to be independent.

Gender-specific PLDA models are trained using the same development data that is used for training LDA transforms. The maximum likelihood estimates of the model parameters $\{\boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_\epsilon\}$ are obtained using an Expectation Maximization (EM) algorithm [15].

E. Hypothesis Testing

Given two length normalized vectors $\mathbf{q}_1$, $\mathbf{q}_2$ of a speaker verification trial, we need to test whether they belong to the same speaker ($\mathcal{H}_s$) or different speakers ($\mathcal{H}_d$). For the Gaussian PLDA above, the log-likelihood ratio can be computed in a closed-form as

$$\begin{aligned} score &= \log \frac{p(\mathbf{q}_1, \mathbf{q}_2|\mathcal{H}_s)}{p(\mathbf{q}_1|\mathcal{H}_d)p(\mathbf{q}_2|\mathcal{H}_d)} \\ &= \log \frac{\mathcal{N}\left(\begin{bmatrix}\mathbf{q}_1\\ \mathbf{q}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\boldsymbol{\Sigma}_\epsilon & \boldsymbol{\Phi}\boldsymbol{\Phi}^T \\ \boldsymbol{\Phi}\boldsymbol{\Phi}^T & \boldsymbol{\Phi}\boldsymbol{\Phi}^T+\boldsymbol{\Sigma}_\epsilon\end{bmatrix}\right)}{\mathcal{N}\left(\begin{bmatrix}\mathbf{q}_1\\ \mathbf{q}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\boldsymbol{\Sigma}_\epsilon & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}\boldsymbol{\Phi}^T+\boldsymbol{\Sigma}_\epsilon\end{bmatrix}\right)}, \end{aligned} \tag{4}$$

where $\mathcal{N}(.; \boldsymbol{\eta}, \boldsymbol{\Lambda})$ is a multivariate Gaussian density with mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{\Lambda}$. The above score can be computed efficiently as described in [13], [17].

TABLE I
DESCRIPTION OF VARIOUS TELEPHONE CONDITIONS OF NIST-08

C6	Telephone speech in training and test
C7	English language telephone speech in training and test
C8	English language telephone speech spoken by a native U.S. English speaker in training and test

TABLE II
EER IN % AND minDCF $\times 10^3$ (SHOWN IN BRACKETS) ON CONDITIONS C6, C7 AND C8 OF NIST-08

System	C6	C7	C8
WLS of AANNs, $\beta = 0$ 240 dim. i-vector (baseline)	12.2 (66.2)	7.2 (38.3)	6.4 (35.6)
WLS of AANNs, back-prop 240 dim. i-vector	10.9 (60.4)	6.4 (31.6)	6.2 (29.6)
Regularized AANNs, $\beta = 0.005$ 240 dim. LDA (proposed)	**10.2** (**56.7**)	**5.4** (**28.1**)	**4.8** (**23.7**)

IV. BASELINE WLS BASED AANN SPEAKER VERIFICATION SYSTEM

The baseline speaker verification system is shown in the Fig. 2. The system has few dissimilarities with the proposed system. The major difference is that the process of adapting the weight matrix of a gender-specific UBM-AANN that impinges on output layer is not regularized. The two different approaches used for adaptation are either to apply back-propagation training as in [6] or to set β to zero in (2). Another difference is that gender dependent low-dimensional subspaces ($\mathbf{T}$ matrices) are trained to capture most of the variability of adapted weights in a WLS sense. Low-dimensional (240) i-vectors are extracted using $\mathbf{T}$ matrices. Subsequently, LDA transforms are learned to rotate the i-vector space. On the other hand, PLDA configuration is same for both the baseline and the proposed systems.

V. EXPERIMENTAL RESULTS

Speaker verification systems are tested on the telephone conditions, described in Table I, of NIST-2008 speaker recognition evaluation (SRE). Table II lists the EER and minimum detection cost function (minDCF) of NIST-2008 for the baseline WLS based AANN (see Section IV) speaker verification system and the proposed regularized AANNs based speaker verification system (see Section III). These neural network systems use the same UBM-AANN of size (39, 20, 6, 39, 39), where each number indicates the number of nodes in a corresponding layer.

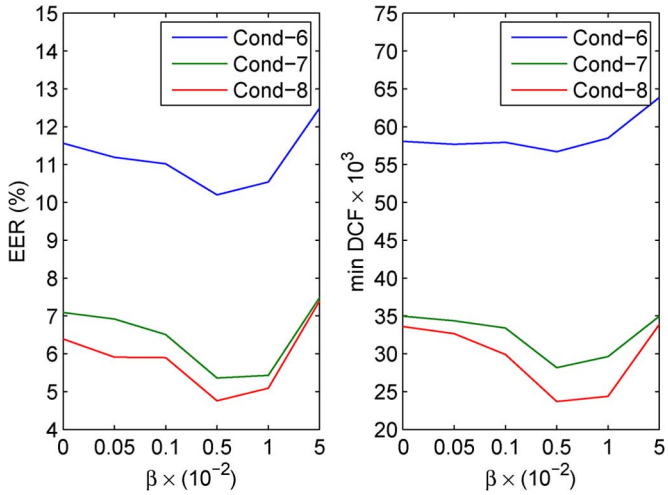

Fig. 4. Error rates of regularized AANNs based speaker verification system on NIST-08 as a function of amount of regularization (β).

The error rates of gender-dependent WLS based AANN speaker verification systems are shown in the first two rows of Table II. These systems project the adapted weight matrices onto 240 dimensional i-vector space and rotate them using LDA. The resultant space is modeled using 140 dimensional (number of columns of $\mathbf{\Phi}$) subspace PLDA model. For baseline system (first row of Table II), β is set to zero in (2) when computing the adapted weight matrices. For the other system (in second row of Table II), back-propagation algorithm[2] is applied for computing the adapted weight matrices. The proposed regularized AANNs based speaker verification system (see Fig. 3) results are listed in third of the table. A non-zero regularization ($\beta = 0.005$) is used in (2) when computing the adapted weight matrices. Subsequently, adapted weights are projected onto a 240 dimensional space using LDA. As in the baseline system, 140 dimensional subspace PLDA model is used for hypothesis testing. It can be observed that the proposed system outperforms the baseline system and yields a relative improvement of 20.9% in EER and 22.5% in minDCF over the baseline.

The effect of changing β on the error rates of the proposed regularized AANNs speaker verification system is shown in Fig. 4. As expected, the system performance increases with the regularization and starts degrading for β greater than 0.005.

VI. Discussion and Conclusions

It was observed that the baseline speaker verification system (see Fig. 2) when the UBM-AANN is adapted with regularization ($\beta = 0.005$), it yields comparable results to the proposed system (see Fig. 3). This result along with the results in Table II indicate that the subspace training ($\mathbf{T}$ matrix) and i-vector extraction steps are unnecessary, and regularized adaptation of UBM-AANN holds the key for obtaining the best performance. These observations suggest a simpler speaker verification system, shown in Fig. 3.

In this paper a closed-form expression for adapting the UBM-AANN with regularization is derived. We have shown that regularized adaptation of UBM-AANN helps improving the speaker verification system performance. Moreover, this also results in a much simpler system.

[2]Back-propagation can be thought of as some form of regularization because the final adapted weights differ from the solution with β set to zero in (2).

Acknowledgment

Authors would like to thank Dr. Daniel Garcia-Romero for helpful discussions.

References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.
[2] B. Yegnanarayana and S. P. Kishore, "Aann: An alternative to gmm for pattern recognition," *Neural Netw.*, vol. 15, no. 3, pp. 459–469, 2002.
[3] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE J.*, vol. 37, no. 2, pp. 233–243, 1991.
[4] I. Shajith, M. Hemant, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *IJCNN*, 1999.
[5] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, 2006.
[6] G. S. V. S. Sivaram, S. Thomas, and H. Hermansky, "Mixture of auto-associative neural networks for speaker verification," in *INTERSPEECH*, 2011.
[7] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1912–1924, 1999.
[8] M. Athineos, H. Hermansky, and D. Ellis, "Plp2 autoregressive modeling of auditory-like 2-d spectrotemporal patterns," in *INTERSPEECH*, 2004.
[9] M. Athineos and D. P. W. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5237–5245, 2007.
[10] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Process. Lett.*, vol. 15, pp. 681–684, 2008.
[11] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature normalization for speaker verification in room reverberation," in *ICASSP*, 2011, pp. 4836–4839.
[12] The ICSI Quicknet Software Package [Online]. Available: http://www.icsi.berkeley.edu/Speech/qn.html
[13] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011.
[14] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *ICASSP*, 2009, pp. 4057–4060.
[15] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV*, 2007.
[16] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, 2010.
[17] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Odyssey*, 2010.