Autoencoder-based unsupervised domain adaptation for speech emotion recognition

Jun Deng, Zixing Zhang, Florian Eyben, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Deng, Jun, Zixing Zhang, Florian Eyben, and Björn Schuller. 2014. "Autoencoder-based unsupervised domain adaptation for speech emotion recognition." *IEEE Signal Processing Letters* 21 (9): 1068–72. https://doi.org/10.1109/lsp.2014.2324759.

Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition

Jun Deng, *Student Member, IEEE*, Zixing Zhang, Florian Eyben, *Member, IEEE*, and Björn Schuller, *Member, IEEE*

***Abstract*—With the availability of speech data obtained from different devices and varied acquisition conditions, we are often faced with scenarios, where the intrinsic discrepancy between the training and the test data has an adverse impact on affective speech analysis. To address this issue, this letter introduces an Adaptive Denoising Autoencoder based on an unsupervised domain adaptation method, where prior knowledge learned from a target set is used to regularize the training on a source set. Our goal is to achieve a matched feature space representation for the target and source sets while ensuring target domain knowledge transfer. The method has been successfully evaluated on the 2009 INTERSPEECH Emotion Challenge's FAU Aibo Emotion Corpus as target corpus and two other publicly available speech emotion corpora as sources. The experimental results show that our method significantly improves over the baseline performance and outperforms related feature domain adaptation methods.**

***Index Terms*—Adaptive denoising autoencoders, domain adaptation, speech emotion recognition.**

I. Introduction

SPEECH emotion recognition aims to automatically predict 'correct' emotional states from acoustic (and/or linguistic) parameters as features using machine learning methods. Many speech emotion recognition engines achieve promising performance only under one common assumption, namely that the training and test data instances are drawn from the same corpus and the same feature space for parametrization is used. However, with speech data obtained from different devices and varied recording conditions, we are often faced with scenarios where such data are typically highly dissimilar in terms of acoustic signal conditions, linguistic content, type of emotion (e. g., acted, elicited, or naturalistic), or the type of labeling scheme used, such as categorical or dimensional labels.

Automatic speech recognition (ASR) is faced with many similar mismatch problems, and the speech community has done a considerable amount of the related work to alleviate this mismatch problem. One major research direction focuses on leveraging auto-associative neural networks to minimize the mismatch problem [1], [2].

One other prominent approach to overcome this 'corpus bias' issue is domain adaptation used when the source domain data has a different distribution than the target domain data, but the task remains the same. In general, domain adaptation techniques are categorized into two classes depending on whether the target domain test data is either partially labeled (semi-supervised) or completely unlabeled (unsupervised). In semi-supervised domain adaptation, correspondences of labeled target data are often used to learn domain transformations [3]. However, unsupervised domain adaptation uses strategies which assume a known class of transformations between the domains, the availability of discriminative features which are common to or invariant across both domains, a latent space where the difference in distribution of source and target data is minimal [4], and a mapping 'path' by which the domain transformation maps the source data onto the target domain [5]. Another popular approach for unsupervised domain adaptation is known as importance weighting. This method has recently been shown to lead to significant improvements in acoustic emotion recognition by Hassan *et al.* : they considered to explicitly compensate for acoustic and speaker differences by employing three transfer learning algorithms [6] (i. e., Kernel Mean Matching (KMM) [7], Unconstrained Least-Squares Importance Fitting (uLSIF) [8], and Kullback-Leibler Importance Estimation Procedure (KLIEP) [9]).

Our work is partially inspired by [10], in which Support Vector Machines (SVMs) are used to learn from the source model $\boldsymbol{w^s}$ by regularizing the distance between the learned model $\boldsymbol{w}$ and $\boldsymbol{w^s}$. Extending this idea to an unsupervised scenario, we propose a novel three-stage data-driven approach in this letter. It is based on adaptive denoising autoencoders which can learn from a source training set with the guidance of a template learned previously from target domain adaptation data, which yields a common representation across training and test domains.

II. Proposed Methodology

A. Denoising Autoencoders

A denoising autoencoder (DAE)–a more recent variant of the basic autoencoder consisting of only one hidden layer–is trained to reconstruct a clean 'repaired' input from a corrupted version [11]. In doing so, the learner must capture the structure of the input distribution in order to reduce the effect of the corruption

Manuscript received March 23, 2014; revised May 09, 2014; accepted May 09, 2014. Date of publication May 16, 2014; date of current version May 21, 2014. This work was supported in part by the China Scholarship Council (CSC), and also in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant 338164 (European Research Council Starting Grant 'iHEARu'). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter K. Willett.

J. Deng, Z. Zhang, and F. Eyben are with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany (e-mail: jun.Deng@tum.de; zixing.zhang@tum.de; eyben@tum.de).

B. Schuller is with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany, and also with the Department of Computing, Imperial College London, London, U.K. (e-mail: bjoern.schuller@imperial.ac.uk).

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

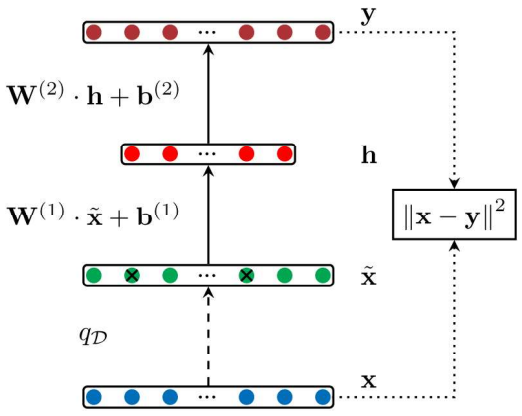

Fig. 1. A denoising autoencoder (DAE) architecture. An input $\mathbf{x}$ is corrupted (via $q_{\mathcal{D}}$) to $\tilde{\mathbf{x}}$. The black crosses ("×") illustrate a corrupted version of the input $\mathbf{x}$ made by $q_{\mathcal{D}}$.

process [12]. It turns out that in this way more robust features are learned compared to a basic autoencoder. The architecture of a DAE is given in Fig. 1.

Formally, an input example $\mathbf{x} \in \mathbf{R}^n$ is first converted to a corrupted version $\tilde{\mathbf{x}}$ by means of a corrupting function $\tilde{\mathbf{x}} \sim q_{\mathcal{D}}(\tilde{\mathbf{x}}|\mathbf{x})$, which could be masking corruption (deleting random elements of the input), additive Gaussian noise, or salt-and-pepper noise in images.

Then, in response to the corrupted example $\tilde{\mathbf{x}}$, the hidden representation $\mathbf{h}(\tilde{\mathbf{x}}) \in \mathbf{R}^m$ is

$$\mathbf{h}(\tilde{\mathbf{x}}) = f(\mathbf{W}^{(1)} \cdot \tilde{\mathbf{x}} + \mathbf{b}^{(1)}), \tag{1}$$

where $f(\cdot)$ is a non-linear activation function, typically a logistic sigmoid function applied component-wise, $\mathbf{W}^{(1)} \in \mathbf{R}^{m \times n}$ is a weight matrix, and $\mathbf{b}^{(1)} \in \mathbf{R}^m$ is a bias vector. It is easily found that the topology structure of the autoencoder completely relies on the size of the input layer n and the number of hidden units m.

The network output maps the hidden representation $\mathbf{h}$ back to a reconstruction $\mathbf{y} \in \mathbf{R}^n$:

$$\mathbf{y} = f(\mathbf{W}^{(2)} \cdot \mathbf{h}(\tilde{\mathbf{x}}) + \mathbf{b}^{(2)}), \tag{2}$$

where $\mathbf{W}^{(2)} \in \mathbf{R}^{n \times m}$ is a weight matrix, and $\mathbf{b}^{(2)} \in \mathbf{R}^n$ is a bias vector.

Given a set of input examples $\mathcal{X}$, the DAE training consists of finding parameters $\theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}\}$ which minimize the reconstruction error. This corresponds to minimizing the following objective function:

$$\mathcal{J}(\theta) = \frac{\lambda}{2}\left(\sum_{l=1}^{2}\sum_{j}\left\|\boldsymbol{w}_j^{(l)}\right\|^2\right) + \sum_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x} - \mathbf{y}\|^2. \tag{3}$$

Here, we also include a weight-decay regularization term with its hyper-parameter λ to the objective function to avoid overfitting. $\boldsymbol{w}_j^{(l)}$ is the j-th column vector of the l-th layer weight matrix $\mathbf{W}^{(l)}$. The minimization is usually realized either by stochastic gradient descent or more advanced optimization techniques such as L-BFGS [13] or conjugate gradient method [14]. In addition, a DAE has an overall asymptotic computational complexity of $O(nm)$ with respect to the network size.

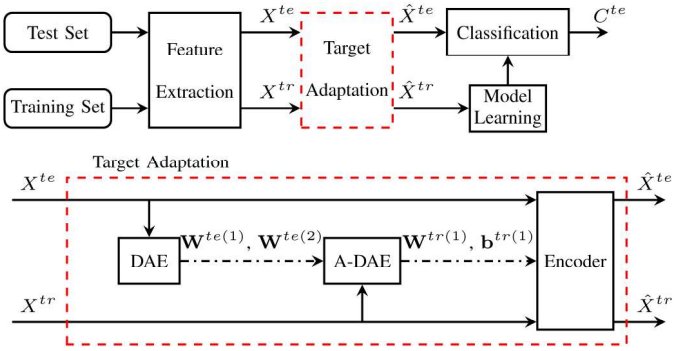

Fig. 2. Overview of the speech emotion recognition system integrating the proposed domain adaptation method. The function "Encoder" refers to the feed-forward procedure (i. e., Eq. (1)) from input data to the activations of the hidden layer of a pre-trained DAE.

B. Learning with Target Prior via Adaptive DAEs

In this letter, we extend the idea for unsupervised domain adaptation by using both a DAE and an adaptive DAE. The aim is to capture source domain knowledge in training an adaptive DAE with the guidance of the prior knowledge previously learned from target domain data by a DAE.

Fig. 2 depicts the affective speech signal analysis method with the proposed domain adaptation method integrated. The proposed method is composed of the following three stages: First, a DAE is learned in a fully unsupervised way from the target domain adaptation data, resulting in the weight matrices $\mathbf{W}^{te(1)}$ (input to hidden layer) and $\mathbf{W}^{te(2)}$ (hidden to output layer) from Eq. (3) as well as the bias vectors $\mathbf{b}^{te(1)}$ and $\mathbf{b}^{te(2)}$.

Next, we propose a new variant of DAEs for domain adaptation, called Adaptive DAE (A-DAE), which force their weights to adapt to the provided weights as well as minimize the reconstruction error between the input and the output at the same time. The output bias vectors $\mathbf{b}^{te}$ of the DAE are not adapted. Hence, given a training example $\mathbf{x} \in \mathcal{X}^{tr}$ and the weights $\mathbf{W}^{te(1)}$ and $\mathbf{W}^{te(2)}$ of a DAE, which were estimated without supervision from the target domain adaptation data (i. e., without knowledge of target labels), the objective function of an A-DAE is formulated as follows:

$$\mathcal{J}^{tr}(\theta) = \frac{\lambda}{2}\left(\sum_{l=1}^{2}\sum_{j}\left\|\boldsymbol{w}_j^{tr(l)} - \beta\boldsymbol{w}_j^{te(l)}\right\|^2\right) + \sum_{\mathbf{x}\in\mathcal{X}^{tr}}\left\|\mathbf{x} - \mathbf{y}^{tr}\right\|^2, \tag{4}$$

where the hyper-parameter β controls the amount of transfer regularization. The weights $\mathbf{W}^{tr(1)}$ and $\mathbf{W}^{tr(2)}$ are initialized randomly and learned during training, while the weights $\mathbf{W}^{te(1)}$ and $\mathbf{W}^{te(2)}$ are kept constant during training.

Without loss of generality, the intuition of the adaptive DAE for domain adaptation can be understood by expanding the weight-decay regularization term:

$$\left\|\boldsymbol{w}^{tr} - \beta\boldsymbol{w}^{te}\right\|^2 = \left\|\boldsymbol{w}^{tr}\right\|^2 + \beta^2\left\|\boldsymbol{w}^{te}\right\|^2 - 2\beta\left\|\boldsymbol{w}^{tr}\right\|\left\|\boldsymbol{w}^{te}\right\|\cos\theta, \tag{5}$$

TABLE I
SUMMARY OF THE THREE CHOSEN AFFECTIVE SPEECH DATABASES

Corpus	Type	Age	Language	Speech	Emotion	# Valence -	# Valence +	# All	#m	#f	Rec	Rate kHz
FAU AEC	target	children	German	variable	natural	3 358 / 2 465	6 601 / 5 792	9 959 / 8 257	21	30	normal	16
ABC	source	adults	German	fixed	acted	213	217	430	4	4	studio	16
SUSAS	source	adults	English	fixed	natural	1 616	1 977	3 593	4	3	noisy	8

Number of instances per binary valence (# Valence, Negative (-), Positive (+)), and overall number (# All)–for FAU AEC divided into official training and test set by "/". Number of female (#f) and male (#m) subjects. Recording conditions (studio/normal/noisy).

where θ is the angle between the two column vectors $\boldsymbol{w}^{tr}$ and $\boldsymbol{w}^{te}$.

On one hand, apart from minimizing the original term $\|\boldsymbol{w}^{tr}\|^2$, the optimization problem aims to use the term $-2\beta\|\boldsymbol{w}^{tr}\|\|\boldsymbol{w}^{te}\|\cos\theta$ to make the transfer by maximizing $\cos\theta$, which is equivalent to minimizing the angle θ between the $\boldsymbol{w}^{tr}$ and $\boldsymbol{w}^{te}$. On the other hand, the term $\|\mathbf{x}-\mathbf{y}^{tr}\|^2$ in the objective function also causes $\|\boldsymbol{w}^{tr}\|$ to adjust to the training data and prevents $\boldsymbol{w}^{tr}$ being close to $\boldsymbol{w}^{te}$. Thus, an adaptive DAE training consists of optimizing a trade-off between the reconstruction error on the training data and target domain knowledge transfer.

Finally, we encode test data and training data via Eq. (1) using the weights ($\mathbf{W}^{tr(1)}$ and $\mathbf{b}^{tr(1)}$) learned by the adaptive DAE. Then, this transformed representation of the training data is used to train a standard supervised classifier (e. g., SVM) for speech emotion recognition as shown in Fig. 2, while the transformed test data is used for evaluation.

III. EXPERIMENTS

A. Selected Data and Acoustic Features

To investigate the performance of the proposed method, we consider the INTERSPEECH 2009 Emotion Challenge two-class task [15]. It is based on the spontaneous FAU Aibo Emotion Corpus (FAU AEC), which contains recordings of 51 children at the age of 10–13 years interacting with the pet robot Aibo in German speech. The details of the challenge's two-class 'negative' versus 'idle' emotions task are given in Table I.

In our experiments, two further publicly available and popular databases, namely the Airplane Behavior Corpus (ABC) [16], and the Speech Under Simulated and Actual Stress (SUSAS) database [17] are chosen as training sets, which are highly different from the FAU AEC in terms of speaker age (adults vs. children in FAU AEC), spoken language (English vs. German in FAU AEC), type of emotion (partially acted vs. naturalistic emotion in FAU AEC), degree of spontaneity and phrase length, type of recording situation, and annotators and subjects. For comparability with FAU AEC, we have to map the diverse emotion classes onto the valence axis of the dimensional emotion model in order to generate a unified set of labels. Binary valence labels according to the mapping shown in Table II are thus generated. Table I summarizes the properties and statistics of the three databases (FAU AEC, ABC, and SUSAS).

TABLE II
MAPPING OF EMOTION CATEGORIES ONTO NEGATIVE AND POSITIVE VALENCE LABELS FOR THE THREE DATABASES

Corpus	Negative	Positive
FAU AEC	negative	idle
ABC	aggressive, intoxicated, nervous, tired	cheerful, neutral, rest
SUSAS	high stress, screaming, fear	medium stress, neutral

For acoustic features, we use a standard set in the field, namely the INTERSPEECH 2009 Emotion Challenge [15] baseline feature set. It consists of 12 functionals applied to 2×16 acoustic Low-Level Descriptors (LLDs) including their first order delta regression coefficients. Thus, the size of the feature vector for each utterance is $16 \times 2 \times 12 = 384$. To ensure reproducibility, the open-source toolkit openSMILE[1] version 2.0 was used with the pre-defined challenge configuration.

B. Experimental Setup

As classifier, we used linear SVMs–as were used in the official baseline of the challenge [15]–with a fixed penalty factor of $C = 0.5$.

For the training of the autoencoders the toolkit minFunc[2] was applied which implements L-BFGS to optimize the parameters of DAEs and A-DAEs. For training of the DAE, we injected masking noise with a variance of 0.01 to generate a corrupted input. For the parameters of the DAE, the weight decay values λ were set to 0.0001, the number of epochs of DAE training was set to 250. In the A-DAE learning process, the hyper-parameter β was fixed to 0.05.

We evaluate the performance of the baseline systems and the A-DAE systems using the evaluation measure of the INTERSPEECH 2009 Emotion Challenge: unweighted average recall (UAR). It is the unweighted average of the per-class recall rates and better reflects overall accuracy in the given case of class imbalance.

C. Comparison to State-of-the-Art Methods

We compare the following methods to evaluate our proposed approach in the context of the current state-of-the-art: (1) Matched Instance Number Training (MINT): randomly (repeated ten times) picks a number of instances from the FAU AEC training set to train an SVM, i. e., without the need of transferring to an intra-corpus scenario. For fair comparison, this number is set by the number of training instances of the

[1]http://sourceforge.net/projects/opensmile/

[2]http://www.di.ens.fr/ mschmidt/Software/minFunc.html

TABLE III
AVERAGE UAR OVER TEN TRIALS: MATCHED INSTANCE NUMBER TRAINING (MINT), CROSS TRAINING (CT), COVARIATE SHIFT ADAPTATION METHODS KLIEP, UNIF, AND KMM, DAE-BASED REPRESENTATION LEARNING, AND THE PROPOSED A-DAE METHOD RELATED TO TRAINING WITH ABC AND SUSAS

UAR [%]	MINT	CT	KLIEP [8]	uLSIF [9]	KMM [7]	DAE [4]	A-DAE
ABC	58.32 ± 4.23	55.28 ± 0.00	55.07 ± 3.81	53.75 ± 1.68	62.52 ± 0.00	55.86 ± 0.80	$\mathbf{64.18 \pm 0.23}$
SUSAS	62.41 ± 3.85	57.32 ± 0.00	58.11 ± 3.56	57.94 ± 0.60	60.43 ± 0.00	62.03 ± 0.69	$\mathbf{62.74 \pm 0.27}$

the ABC or SUSAS sets, respectively. (2) Cross Training (CT): uses ABC or SUSAS to train the standard (SVM) classifier, which is the 'classical' cross-corpus testing, i. e., it involves no adaptation. (3) KLIEP [8], (4) uLSIF [9], and (5) KMM [7]: utilize the modern domain adaption methods on the ABC and SUSAS database for covariate shift adaptation, respectively. We choose the 'tuning parameters' following [6]. (6) DAE: employs denoising autoencoders for representation learning in order to match training examples to test examples; this was successfully applied to the transfer learning challenge and domain adaptation [4], [18].

We study the cross-corpus setting with the number of hidden units fixed to 256 where we train acoustic emotion recognition models on ABC or SUSAS while evaluating on the FAU AEC test set (except the MINT condition that uses FAU AEC data for training). We report results of the averaged UAR over the ten trials in Table III. As can be seen, our approach always shows a comparable performance to other approaches [4], [7]–[9].

For the small database ABC, the two standard methods (CT and MINT) only yield an average UAR around chance level (55.28% and 58.32%). With the benefits of compensation for the existent mismatch, the covariate shift adaptation KMM can achieve the accuracy of 62.52%. The proposed A-DAE method outperforms all other methods with 64.18% UAR. This improvement has a statistical significance at $p < 0.001$ with a one-sided z-test when compared to CT and MINT.

On the SUSAS database, our proposed method shows a significant improvement over other methods. Specifically, the A-DAE method gives an average UAR of 62.74%, which is slightly higher than the maximum average UAR obtained by MINT. Moreover, it passes the significance test at $p < 0.001$ and $p < 0.002$ against the CT and KMM methods, respectively. In the mean time, it is worth noting that the average UAR obtained by MINT increases dramatically to 62.42% just due to the larger size of SUSAS leading to more instances being chosen from the FAU AEC training set in comparison to ABC.

D. A-DAE vs. DAE

We now compare the A-DAE and DAE methods in detail. In Fig. 3, we provide UAR for different numbers of hidden units m, where we observe performance changes for different parameter settings. Based on Fig. 3, it is worth noting that the proposed method obtains the highest UAR of 64.67% for ABC and of 63.02% for SUSAS at $m = 1024$ and $m = 512$, respectively. Surprisingly, we could not obtain a sustained performance growth with more hidden units for SUSAS. One reason is that the utterances of ABC are more complex and have more variance (length and content) than those of SUSAS which contain pre-defined short commands. Therefore, the increase in hidden units potentially yields to more generalization performance for ABC than for SUSAS. In contrast, increasing the number of hidden units to $m = 1024$ in the case of SUSAS reduces the corresponding performance because overfitting occurs. Nevertheless, increasing the number of hidden units leads to additional improvement indeed, which confirms that an over-complete first hidden layer works better than an under-complete one when using unsupervised pre-training as in the theory of deep architectures [19].

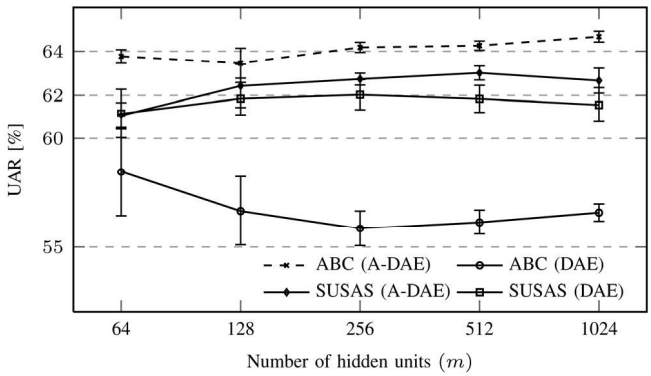

Fig. 3. Average UAR with standard deviation over ten trials with varying number of hidden units (m) using DAE or A-DAE.

IV. CONCLUSIONS

In this letter, we proposed a novel unsupervised domain adaptation method based on adaptive denoising autoencoders for affective speech signal analysis. The method is capable of reducing the discrepancy between training and test sets due to different conditions (e. g., different corpora). We first built a denoising autoencoder on the target domain adaptation set without using any label information with the aim to encode the target data in an optimal way. These encoding parameters are used as prior information to regularize the training process of an A-DAE on the training set. In this way, a trade-off between the reconstruction error on the training data and a knowledge transfer to the target domain is found, effectively reducing the existing mismatch between the training and testing conditions in an unsupervised way. Results with three publicly available corpora show that the proposed method effectively and significantly enhances the emotion classification accuracy in mismatched training and test conditions when compared to other domain adaptation methods. In future work, we plan to use the dropout strategy [20] to further improve the generalization performance of autoencoder-based domain adaptation and extend A-DAEs to deep architectures.

REFERENCES

[1] S. Garimella, S. Mallidi, and H. Hermansky, "Regularized auto-associative neural networks for speaker verification," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 841–844, Dec. 2012.

[2] S. P. Kishore and B. Yegnanarayana, "Speaker verification: Minimizing the channel effects using autoassociative neural network models," in *Proc. ICASSP*, Istanbul, Turkey, 2000, vol. 2, pp. 1101–1104.

[3] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. ACII*, Geneva, Switzerland, 2013, pp. 511–516.

[4] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. ICML*, Bellevue, WA, USA, 2011, pp. 513–520.

[5] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Patt. Anal. Mach. Intell.*, 2014, to be published.

[6] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1458–1468, 2013.

[7] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset Shift Mach. Learn.*, vol. 3, no. 4, pp. 131–160, 2009.

[8] T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," in *Proc. NIPS*, Vancouver, BC, Canada, 2008, pp. 809–816.

[9] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation with model selectionand its application to covariate shift adaptation," in *Proc. NIPS*, Vancouver, BC, Canada, 2007, pp. 1433–1440.

[10] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 2252–2259.

[11] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, Helsinki, Finland, 2008, pp. 1096–1103.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Math. Progr.*, vol. 45, no. 1–3, pp. 503–528, 1989.

[14] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems1," *J. Res. Nat. Bur. Stand.*, vol. 49, no. 6, 1952.

[15] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009Emotion Challenge," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 312–315.

[16] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. ICASSP*, Honolulu, HI, USA, 2007, pp. 733–736.

[17] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1743–46.

[18] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML*, Bellevue, WA, USA, 2011, pp. 17–36.

[19] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade.* Berlin, Germany: Springer, 2012, pp. 437–478.

[20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580 2012.