Universum autoencoder-based domain adaptation for speech emotion recognition

Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Fruhholz, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Deng, Jun, Xinzhou Xu, Zixing Zhang, Sascha Fruhholz, and Björn Schuller. 2017. "Universum autoencoder-based domain adaptation for speech emotion recognition." In *IEEE Signal Processing Letters*, 24:500–504. New York, NY: Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/lsp.2017.2672753.

Universum Autoencoder-Based Domain Adaptation for Speech Emotion Recognition

Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller, *Senior Member, IEEE*

***Abstract*—One of the serious obstacles to the applications of speech emotion recognition systems in real-life settings is the lack of generalization of the emotion classifiers. Many recognition systems often present a dramatic drop in performance when tested on speech data obtained from different speakers, acoustic environments, linguistic content, and domain conditions. In this letter, we propose a novel unsupervised domain adaptation model, called Universum autoencoders, to improve the performance of the systems evaluated in mismatched training and test conditions. To address the mismatch, our proposed model not only learns discriminative information from labeled data, but also learns to incorporate the prior knowledge from unlabeled data into the learning. Experimental results on the labeled Geneva Whispered Emotion Corpus database plus other three unlabeled databases demonstrate the effectiveness of the proposed method when compared to other domain adaptation methods.**

***Index Terms*—Deep learning, domain adaptation, speech emotion recognition, universum autoencoders ($\mathfrak{U}$-AE).**

I. Introduction

DESPITE remarkable advances in speech emotion recognition [1]–[3], robustness and generalization of emotion recognition systems remain an open challenge [4]–[6]. The difficulty of the task arises from the enormous speech variability that an emotion recognition system confronts when it is tested with new speakers and environments. While it may be possible to gain access to all variations by collecting and annotating large amounts of emotional speech data that might be presented to the system, such approaches are infamously time consuming and expensive, and tend to fail when the target domain changes.

An appealing approach to enhancing the generalization in speech emotion recognition is domain adaptation. In such systems, the mismatch between the training and test data is addressed by leveraging over prior knowledge found in one source. Previous studies have proposed a variant of domain adaptation approaches with positive performance for speech emotion recognition, including kernel mean matching (KMM) [7], a Kullback–Leibler importance estimation procedure (KLIEP) [8], shared-hidden-layer autoencoders (SHLA) [9], Nonnegative matrix factorization [10], domain adaptive least-squares regression [11], and PCANet [12]. These domain adaptation methods usually adopt a hybrid framework that combines features of unsupervised learning (e. g., autoencoders) and supervised learning [e. g., support vector machines (SVM)]. The critical disadvantage of these existing methods is that the unsupervised learning is prone to learn irrelevant representations making the following supervised classifier unfavorable to its learning, due to the absence of the label information of the task of interest.

In this letter, we propose a novel end-to-end domain adaptation algorithm, called *Universum autoencoder* ($\mathfrak{U}$-AE), which endows an unsupervised learning autoencoder with the supervised learning capability. Such additional supervised learning capability enables this novel autoencoder to guide the learning toward the aim to improve the performance of systems. Our approach is motivated by the observation that each emotionally colored speech audio will contain common knowledge of emotions that are similar to those in target speech audios. Therefore, if we can exploit such knowledge from the unlabeled data, then this knowledge can be incorporated into mitigating the mismatch.

Our work is theoretically inspired by Universum learning ($\mathfrak{U}$-learning) [13], [14], where unlabeled data, termed as Universum, were used as a penalty term of the standard SVM objective function to enhance classification performance for digit recognition. Extending this idea to a domain adaptation scenario, we propose to add the margin-based loss introduced in $\mathfrak{U}$-learning to a deep autoencoder, leading to reducing the inherent mismatch between the training and test data by simultaneously learning common knowledge from labeled and unlabeled data.

II. Universum Autoencoders

In $\mathfrak{U}$-AE, we are given a labeled training set of N examples$\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\ldots,(\mathbf{x}_N,\mathbf{y}_N)\}$, where $\mathbf{y}_i \in$

This work was in part supported by the BMBF IKT2020-Grant under Grant 16SV7213 (EmotAsS) and in part by the European Community's Seventh Framework Programme through the ERC Starting under Grant 338164 (iHEARu). The work of S. Frühholz was supported by the Swiss National Science Foundation under Grant PP00P1_157409/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter K. Willett. *(Corresponding author: Jun Deng.)*

J. Deng and Z. Zhang are with the Chair of Complex and Intelligent Systems, University of Passau, Passau 94032, Germany (e-mail: jun.deng@uni-passau.de; zixing.zhang@uni-passau.de).

X. Xu is with the Machine Intelligence and Signal Processing Group, Technische Universität München, Munich 80333, Germany (e-mail: xinzhou.xu@tum.de).

S. Frühholz is with the Institute of Psychology, University of Zurich, Zurich 8006, Switzerland, the Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich 8092, Switzerland, and also with Center for Integrative Human Physiology, University of Zurich, Zurich 8006, Switzerland (e-mail: sascha.fruehholz@psychologie.uzh.ch).

B. Schuller is with the Chair of Complex and Intelligent Systems, University of Passau, Passau 94032, Germany, and also with the Department of Computing, Imperial College London, London, U.K. (e-mail: schuller@tum.de).

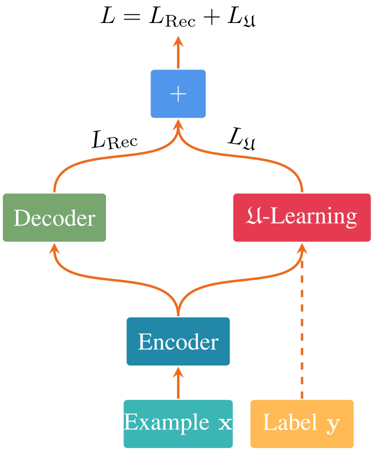


Fig. 1. $\mathfrak{U}$-AE consist of the encoder, the decoder, and the $\mathfrak{U}$-learning path. Given a labeled training set of N examples$\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ and an unlabeled dataset, called the Universum $\mathfrak{U} = \{\mathbf{x}^\star_{N+1}, \mathbf{x}^\star_{N+2}, \ldots, \mathbf{x}^\star_{N+M}\}$, the objective function L jointly minimizes the reconstruction error L_{Rec} and the $\mathfrak{U}$-learning loss $L_{\mathfrak{U}}$ in the form of a combination of the L_2-margin loss and ϵ-insensitive loss.

$\mathbb{R}^{K\times 1}$, $\mathbf{y}_i[k] = 1$ when sample i belongs to class k, otherwise $\mathbf{y}_i[k] = 0$, and K is the total number of classes. In addition, we are given an unlabeled dataset, called the Universum $\mathfrak{U} = \{\mathbf{x}^\star_{N+1}, \mathbf{x}^\star_{N+2}, \ldots, \mathbf{x}^\star_{N+M}\}$. Clearly, as in domain adaptation, we simply assume that the unlabeled data should be potentially relevant to the test data if unlabeled data are to provide prior knowledge. For example, we would normally expect that the unlabeled data and the test data come from the same input spoken language type (e. g., German language) or speech production modality (e. g., “normal phonated” speech), etc.

A. Encoder and the Decoder

As illustrated in Fig. 1, the $\mathfrak{U}$-AE model includes the *encoder* and the*decoder*, which both consist of multiple feed-forward neural network layers, resulting in an algorithm that has the representational capability to discover intricate structures in the input data in an unsupervised manner. The encoder takes the input $\mathbf{x}$ and nonlinearly maps it to a hidden representation $\mathbf{h}^L_e$

$$\mathbf{h}^L_e\left(\mathbf{W}^e, \mathbf{x}\right) = f\left(\mathbf{W}^e_L f\left(\mathbf{W}^e_{L-1} \cdots f\left(\mathbf{W}^e_1 \mathbf{x}\right) \cdots\right)\right) \quad (1)$$

where the matrix $\mathbf{W}^e_l$ corresponds to the parameters of the lth layer, $f(\cdot)$ is a nonlinear activation function, such as the rectified linear unit in this letter, and L represents the number of layers.

The decoder is required to reconstruct the input from the resulting representation $\mathbf{h}^L_e$ via multiple nonlinear processing layers

$$\hat{\mathbf{x}}\left(\mathbf{W}^d, \mathbf{h}^L_e\right) = \mathbf{W}^d_L f\left(\mathbf{W}^d_{L-1} \cdots f\left(\mathbf{W}^d_1 \mathbf{h}^L_e\right) \cdots\right). \quad (2)$$

We define the *reconstruction error* as the *sum of squared error* within all of the labeled data and unlabeled data

$$L_{\text{Rec}}\left(\mathbf{W}\right) = \sum_{i=1}^{N+M} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \quad (3)$$

whose value is expected to be minimized during training. The parameters $\mathbf{W}$ are determined by minimizing $L_{\text{Rec}}(\mathbf{W})$.

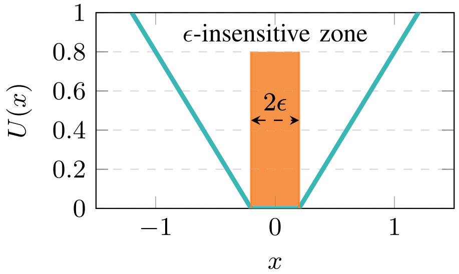


Fig. 2. ϵ-insensitive loss function $(U(x) = \max(|x| - \epsilon, 0))$ penalizes Universum examples that are far from the ϵ-insensitive zone. Here, it is shown with $\epsilon = 0.2$.

B. Universum Learning Path

In addition to the aforementioned reconstruction process, the $\mathfrak{U}$-AE is equipped with the$\mathfrak{U}$-*learning path*, leading the autoencoder to efficiently learning from the labeled and unlabeled data. The $\mathfrak{U}$-learning path derives from a $L_{\mathfrak{u}}$-layer feed-forward neural network on top of the outputs of the encoder [i. e., (1)]:

$$\mathbf{h}^{\mathfrak{u}}\left(\mathbf{W}^{\mathfrak{u}}, \mathbf{h}^L_e\right) = \mathbf{W}^{\mathfrak{u}}_{L_{\mathfrak{u}}} f\left(\mathbf{W}^{\mathfrak{u}}_{L_{\mathfrak{u}}-1} \cdots f\left(\mathbf{W}^{\mathfrak{u}}_1 \mathbf{h}^L_e\right) \cdots\right) \quad (4)$$

where $\mathbf{h}^{\mathfrak{u}} \in \mathbb{R}^{K\times 1}$ represents the outputs from the last layer of the $\mathfrak{U}$-learning path.

We apply a novel loss function to optimize the parameters in the $\mathfrak{U}$-AE, which is originally used in $\mathfrak{U}$-SVM [14]. The loss function is defined as the mixture of the L_2-margin loss and ϵ-insensitive loss. Here, the aim is to take advantage of the labeled and unlabeled data.

For the labeled data, the L_2 margin-based loss is formulated as follows:

$$L_2 = C \sum_{i=1}^{N} \max(\xi - \mathbf{y}_i^T \mathbf{h}^{\mathfrak{u}}_i, 0)^2 \quad (5)$$

where the margin value ξ is a tuning-parameter, and $C > 0$ is a regularization parameter controlling the effect of margin.

In neural networks, the weights are commonly learnt by the backpropagation algorithm [15]. Hence, we need the gradients from the L_2 margin-based loss with respect to its input, which is calculated as follows:

$$\frac{dL_2}{d\mathbf{h}^{\mathfrak{u}}_i} = -2C \max(\xi - \mathbf{y}_i^T \mathbf{h}^{\mathfrak{u}}_i, 0). \quad (6)$$

For the unlabeled data, we adopt the ϵ-insensitive loss, depicted in Fig. 2, which is defined as follows:

$$\begin{aligned} U(\mathbf{h}^{\mathfrak{u},\star}) &= H_{-\epsilon}(\mathbf{h}^{\mathfrak{u},\star}) + H_{-\epsilon}(-\mathbf{h})^{\mathfrak{u},\star} \\ &= H_{-\epsilon}(|\mathbf{h}^{\mathfrak{u},\star}|) = \sum_{k=1}^{K} \mathtt{max}(|\mathbf{h}^{\mathfrak{u},\star}[k]| - \epsilon, 0) \quad (7) \end{aligned}$$

where the Hinge loss function $H_{-\epsilon}(\mathbf{h}^{\mathfrak{u},\star}) = \sum_{k=1}^{K} \mathtt{max}(\mathbf{h}^{\mathfrak{u},\star}[k] - \epsilon, 0)$, and ϵ is a tuning-parameter. This loss measures the real-valued output of our network on the Universum examples and penalizes outputs that are far away from zero. That is, we wish to minimize the total loss

$$L_{\mathfrak{u}} = \sum_{i=N+1}^{N+M} U(\mathbf{h}^{\mathfrak{u},\star}_i). \quad (8)$$

From the theory of $\mathfrak{U}$-learning, the generalization of a classifier can be enhanced by learning through contradictions [14]. Thus, the intuition behind the ϵ-insensitive loss is to find an equivalence class with a large number of contradictions on the Universum.

We add the ϵ-insensitive loss as a penalty term to the L_2 margin-based loss function to form the total loss function of the $\mathfrak{U}$-learning path. That is, we minimize

$$L_{\mathfrak{U}} = CL_2 + C_{\mathrm{u}} L_{\mathrm{u}} \tag{9}$$

where the tuning-parameters $C > 0$ and $C_{\mathrm{u}} > 0$ control the tradeoff between minimization of classification errors on the labeled data and maximization of the number of contradictions on the Universum data.

C. Joint Objective

The final objective function is a convex linear combination of the reconstruction error [i. e., (3)] and the margin-based loss [i. e., (9)]

$$L = L_{\mathrm{Rec}} + CL_2 + C_{\mathrm{u}} L_{\mathrm{u}}. \tag{10}$$

In this way, the $\mathfrak{U}$-AE makes full use of the available labeled and unlabeled data (Universum) to learn a strong classifier with a minimum reconstruction error as well as a minimum classification error. Like the standard neural networks, the present model can be optimized by the backpropagation algorithm.

Analogous to denoising autoencoders (see [5], [9], [16]), we intentionally inject noise, such as additive Gaussian noise, into the inputs so as to encourage the $\mathfrak{U}$-AE to learn meaningful representations. Furthermore, *batch normalization* [17] is applied for addressing the internal covariance shift issue.

III. Experiments

A. Selected Data and Acoustic Features

To investigate the performance of the proposed method, we consider a typical cross-domain speech emotion recognition task, namely the cross-speech-mode task introduced as in the Geneva Whispered Emotion Corpus (GeWEC) [18]. The selected corpus provides *normal* and *whispered* paired utterances. Two male and two female professional French-speaking actors in Geneva were recruited to speak eight predefined French pseudowords with a given emotional state in both normal and whispered speech modes. Speech was expressed in four emotional states: *angry*, *frightened*, *happy*, and *neutral*. As a result, GeWEC consists of 1280 instances in total. In our experiments, the whispered speech GeWEC data are used for training while the normal speech mode data are used for testing. In this setting, the high mismatch between the training and test data appears to lead a conventional emotion system to poor performance.

In addition, three publicly available and popular databases, namely the *Airplane Behavior Corpus* (ABC) [19], the *Berlin EMOtional speech database* (EMODB) [20], and the *Speech Under Simulated and Actual Stress* (SUSAS) set [21] are chosen as unlabeled training sets. Table I summarizes the properties and statistics of the four databases (GeWEC, ABC, EMODB, and SUSAS).

For acoustic features, we adopted the standardized feature set used in the INTERSPEECH 2009 Emotion Challenge [22], in which 12 functionals are applied to 2×16 acoustic low-level descriptors including their first-order delta regression coefficients. Thus, the total feature vector per utterance contains $16 \times 2 \times 12 = 384$ attributes. To ensure reproducibility, the open source openSMILE toolkit version 2.0 [23] was used with the predefined challenge configuration.

B. Experimental Setup

In the neural network learning process, we applied the Adam optimization algorithm [24] with maximum 50 epochs to optimize the parameters. For training the $\mathfrak{U}$-AE neural networks, we injected Gaussian noise with a variance of 0.5 to generate the corrupted input. We used grid search to search over the learning rate $\{0.1, 0.01, 0.001, 0.0001\}$, the number of hidden nodes $\{100, 200, 500, 1\,000\}$, $C \in \{0.01, 0.005, 0.001, 0.00001\}$, $\epsilon \in \{0.005, 0.001\}$, and $C_{\mathrm{u}} \in \{0.005, 0.001, 0.0001\}$. The margin value ξ was fixed to 1 in the experiments. The number of hidden layers for the encoder and decoder is set to two while the $\mathfrak{U}$-learning path has only one hidden layer. Each hidden layer has the same hidden nodes. Input and target features are standardized to zero mean and unit variance on the training set. We apply five-fold validation on the training set to parameter tuning. Each experiment is repeated ten times with different seeds for parameter installation and data selection.

We evaluate the performance by unweighted average recall (UAR), which is often used as the challenges-recommended measure for speech emotion recognition. In addition, significance tests are conducted by computing a one-sided z-test.

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

We compare our $\mathfrak{U}$-AE with the following methods.

1) SVM: A widely adopted system based on the chosen acoustic features and the SVM classifier.
2) KLIEP [8] and
3) KMM [7]: Two representative covariate shift adaptation methods, which were recently shown to yield impressive performance improvement on the first speech emotion challenge task [26].
4) SHLA [9]: A successful autoencoder-based domain adaptation method for emotion recognition.
5) MGD [18]: A recently published model based on the modified group delay (MGD) features that provides the state-of-the-art performance on the GeWEC test set.

Table II summarizes the cross-domain recognition accuracy of all algorithms over ten independent runs. As can be seen, the previous domain adaptation algorithms such as KLIEP obtain an increase in performance on the cross-domain task when compared to the traditional methods such as SVM. This reveals that domain adaptation can reduce the domain mismatch and improve the generalization capacity of the emotion classifier [9]. Notably, our proposed $\mathfrak{U}$-AE method, which simultaneously exploits the labeled training data and the given unlabeled data (e. g., ABC), performs better than all other methods. Specifically, the $\mathfrak{U}$-AE method obtains an average UAR of 63.3 % given the unlabeled ABC data, which is higher than the maximum average UAR (58.3 %) obtained by SHLA. It passes the

TABLE I
SUMMARY OF THE FOUR CHOSEN DATABASES

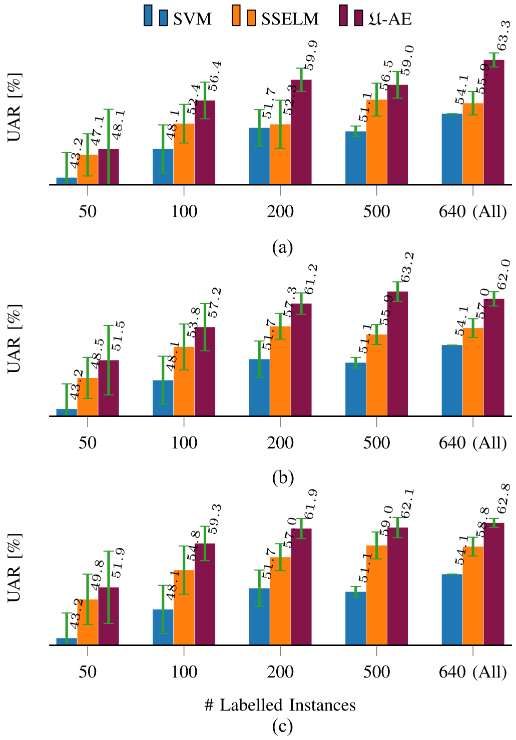
Corpus	Mode	Language	Type	# Emotions	# All	h:mm	#m	#f	Rec	Rate (kHz)
GeWEC	whispered/normal	French	acted	4	1 200	0:14	2	2	studio	16
ABC	normal	German	acted	6	430	1:15	4	4	studio	16
EMODB	normal	German	acted	7	494	0:21	5	5	studio	16
SUSAS	normal	English	natural	4	3 593	1:01	4	3	noisy	8

Mode whispered/normal (speech production). Type of material (acted/natural). Number of emotion categories (# Emotions). Overall number of turns (# All). Total audio time (h:mm). Number of female (# f) and male (# m) subjects. Recording (Rec) conditions (studio/noisy).

TABLE II
AVERAGE UAR OVER TEN TRIALS ON THE GEWEC TEST SET

Method	UAR [%]
Supervised methods:	
SVM	$54.1_{\pm 0.0}$
MGD [18]	$54.8_{\pm 0.0}$
Domain Adaptation methods:	
KLIEP [25]	$56.7_{\pm 0.6}$
KMM [7]	$55.8_{\pm 0.0}$
SHLA [9]	$58.3_{\pm 1.8}$
Our proposed method:	
𝔘-AE (+ABC)	$\mathbf{63.3}_{\pm 1.2}$
𝔘-AE (+EMODB)	$\mathbf{62.0}_{\pm 1.2}$
𝔘-AE (+SUSAS)	$\mathbf{62.8}_{\pm 0.8}$

Unlabeled data are from either ABC, EMODB, or SUSAS. We compare our proposed method with previously reported GeWEC test UARs and other domain adaptation methods. Significant results are bolded. Names of unlabeled data are in parentheses.



(a)

(b)

(c)

Fig. 3. Effect of the number of the labeled training data when different unlabeled data (i. e., ABC, EMODB, or SUSAS) are used for training. The normal supervised learning SVM and a semisupervised learning algorithm SSELM [27] are used for comparison. (a) ABC. (b) EMODB. (c) SUSAS.

significance test at $p < 0.05$ and $p < 0.001$ against the SHLA and MGD methods, respectively. These experimental results indicate the effectiveness of using the concept of the simultaneous reconstruction and 𝔘-learning for domain adaptation.

D. Performance Effect of Different Amount of Labeled Data

We further analyze the performance effect of using different numbers of labeled training data. For comparison, the predominant supervised learning emotion recognition approach by SVM is considered. Further, we believe that we should compare with semisupervised learning algorithms that exploit information in the unlabeled data. For this purpose, we compare with semisupervised extreme learning machines (SSELM) [27]. In the analysis, we randomly selected a fixed number of labeled data from the GeWEC whispered speech, and used them together with one of the three unlabeled databases for the training. Fig. 3 presents the experimental results.

From Fig. 3, we can find that increasing the number of the labeled data leads to a consistent performance growth. The unlabeled data, as expected, help the SSELM and 𝔘-AE methods effectively improve the performance. Surprisingly, SSELM and 𝔘-AE make use of only 100 labeled training examples together with the unlabeled SUSAS data to obtain the UAR of 54.8 % and 59.3 %, respectively, which are even comparable to the SVM UAR 54.1 % with the whole labeled data (640 examples). It can be further observed that 𝔘-AE outperforms SSELM by a margin. This suggests that 𝔘-AE is able to achieve a more efficient use of unlabeled data than the previous semisupervised learning method in the context of domain adaptation.

IV. CONCLUSION

In this letter, we proposed Universum autoencoders, a novel model for unsupervised domain adaptation in speech emotion recognition and general related machine learning tasks. The proposed autoencoder retains the representational capability to discover the intrinsic structures in the input. Furthermore, based on the concept of 𝔘-learning, the novel model leverages the margin-based classification loss to exploit the prior knowledge from unlabeled data in an attempt to regulate the learning process. Cross-speech-mode experiments on the GeWEC data present that the proposed method effectively and significantly enhances the emotion recognition accuracy in mismatched conditions when compared to other domain adaptation methods. In future work, we plan to extend the proposed method to various computational paralinguistics and further tasks.

References

[1] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, Nov. 2013.

[3] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, pp. 1–23, 2012.

[4] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interact.*, Geneva, Switzerland, 2013, pp. 511–516.

[5] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[6] M. Abdel-Wahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brisbane, QLD, Australia, 2015, pp. 5058–5062.

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

[8] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selectionand its application to covariate shift adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2007, pp. 1433–1440.

[9] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Florence, Italy, 2014, pp. 4851–4855.

[10] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, "Speech emotion recognition using transfer non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, 2016, pp. 5180–5184.

[11] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 585–589, May 2016.

[12] Z. Huang, W. Xue, Q. Mao, and Y. Zhan, "Unsupervised domain adaptation for speech emotion recognition using PCANet," *Multimedia Tools Appl.*, pp. 1–15, 2016.

[13] V. N. Vapnik and V. Vapnik, *Statistical Learning Theory*, vol. 1, New York, NY, USA: Wiley, 1998.

[14] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, "Inference with the Universum," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pennsylvania, USA, 2006, pp. 1009–1016.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[16] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 1096–1103.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 448–456.

[18] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Proc. 7th Int. Workshop Spoken Dialogue Syst.*, Saariselkä, Finland, Jan. 2016, p. 6.

[19] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Honolulu, HI, USA, 2007, pp. 733–736.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1517–1520.

[21] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, Rhodes, Greece, 1997, pp. 1743–1746.

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

[23] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[24] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.

[25] T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," in *Proc. 22nd Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2008, pp. 809–816.

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

[27] G. Huang, S. Song, J. N. D. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning machines," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2405–2417, Dec. 2014.