Recognizing Emotions From Whispered Speech Based on Acoustic Feature Transfer Learning

JUN DENG[1], SASCHA FRÜHHOLZ[2,3,4,5], ZIXING ZHANG[1], (Member, IEEE), AND BJÖRN SCHULLER[1,6], (Senior Member, IEEE)

[1]Chair of Complex and Intelligent Systems, University of Passau, 94032 Passau, Germany
[2]Institute of Psychology, University of Zurich, 8006 Zürich, Switzerland
[3]Neuroscience Center Zurich, University of Zurich, 8006 Zürich, Switzerland
[4]ETH Zürich, 8092 Zürich, Switzerland
[5]Center for Integrative Human Physiology, University of Zurich, 8006 Zürich, Switzerland
[6]Department of Computing, Imperial College London, London SW7 2AZ, U.K.

Corresponding author: J. Deng (jun.deng@uni-passau.de)

This work was supported in part by the Bundesministerium für Bildung und Forschung within IKT 2020 through the Emotionssensitives Assistenzsystem for Mensches met Behinderungen Project under Grant 16SV7213, in part by the European Community's Seventh Framework Programme within the European Research Council Starting Grant through the iHEARu Project under Grant 338164, and in part by the European Union's Horizon 2020 Programme within the Research and Innovation Action through the Multi-Modal Human-Robot Interaction for Teaching and Expanding Social Imagination in Autistic Children Project under Grant 688835. The work of S. Frühholz was supported by the Swiss National Science Foundation under Grant PP00P1 157409/1.

ABSTRACT Whispered speech, as an alternative speaking style for normal phonated (non-whispered) speech, has received little attention in speech emotion recognition. Currently, speech emotion recognition systems are exclusively designed to process normal phonated speech and can result in significantly degraded performance on whispered speech because of the fundamental differences between normal phonated speech and whispered speech in vocal excitation and vocal tract function. This study, motivated by the recent successes of feature transfer learning, sheds some light on this topic by proposing three feature transfer learning methods based on denoising autoencoders, shared-hidden-layer autoencoders, and extreme learning machines autoencoders. Without the availability of labeled whispered speech data in the training phase, in turn, the three proposed methods can help modern emotion recognition models trained on normal phonated speech to reliably handle also whispered speech. Throughout extensive experiments on the Geneva Whispered Emotion Corpus and the Berlin Emotional Speech Database, we compare our methods to alternative methods reported to perform well for a wide range of speech emotion recognition tasks and find that the proposed methods provide significant superior performance on both normal phonated and whispered speech.

INDEX TERMS Speech emotion recognition, whispered speech, feature transfer learning, autoencoders, extreme learning machines.

I. INTRODUCTION

Speech emotion recognition has grown into a major research topic in speech processing, human-computer interaction, and computer-mediated human communication over the last decades (see [1]–[4]). In general, it focuses on using machine learning methods to automatically predict 'correct' emotional states from speech. Apart from normal phonated speech at which current studies mainly have made considerable efforts to date, in fact, *whispered speech* is another common form of speaking to communicate, which is produced by speaking with high breathiness and no periodic excitation. With the absence of periodic vibration of the vocal folds during the production, whispered speech structure is significantly altered which results in reduced perceptibly and a significant reduction in intelligibility. In the meantime, it was already found that whispered speech can encode prosodic information, and thereby still convey clues carrying emotion information [5], [6]. Naturally, whispered speech plays an important role in our daily life in order to intentionally confine the hearing of speech to listeners who are nearby. For example, we whisper to the user interface over the cellphone when offering privacy information in terms of date of birth, credit card information, billing address to make hotel, flight, and table reservations. Another

area of interest are patients with speech disabilities who are affected by a temporary or long-term in the vocal fold structure or disease of the vocal system such as functional aphonia or laryngeal disorders [7] and therefore can only produce whisper-like sounds. For speech emotion recognition, however, only a handful of efforts have been devoted to recognizing whispered speech by now (i. e., [8], [9]). Especially, the issue of how to build a practically feasible emotion recognition system for whispered speech has not been addressed yet, as past work mainly analyzed the differences of the prosodic features in emotions of Chinese whispered speech [8], [9]. Hence, to be more useful in practice it would be highly desirable to enable an emotion recognition system to process whispered speech as well with promising accuracy.

In the speech community, there has been a considerable amount of the related work on whispered speech [7], [10]–[17]. In [14], Janke *et al.* addressed the F0 modeling in whisper-to-audible speech conversion and then proposed a hybrid unit selection approach for whisper-to-speech conversion based on the finding that F0 contours can be derived from the mapped spectral vectors. Furthermore, to improve the intelligibility of whispered speech in various noise contexts, an unsupervised learning of phonemes was proposed based on convolutive non-negative matrix factorization [12]. For the task of acoustic voice analysis in computer laryngeal diagnostic, Mitev and Hadjitodorov [7] developed three methods for fundamental frequency determination of voice of patients with laryngeal disorders, including autocorrelation method, spectral method, and cepstral method. Moreover, these methods were combined in a system for acoustic analysis and screening of the pathological voices in the everyday clinical practice.

Rather than tediously collecting and labeling whispered speech and designing a dedicated system from scratch, past studies also have shown that a workable scheme in an attempt to deal with whispered speech is to explore normal phonated speech data to create and develop systems that would be much more robust against variability and shifts in speech modes (e. g., normal phonated and whispered modes) [10], [18]. For example, Fan and Hansen [10] recently considered a feature transformation estimation method in the training phase which results in a more robust speaker model for speaker identification on whispered speech. Three estimation methods are proposed to model the transformation from normal phonated speech to whispered speech. This solution seems also reasonably feasible and worthwhile in speech emotion recognition, because it allows for one single recognition system to process both normal phonated and whispered speech simultaneously. Another important reason is that massively available normal phonated speech is a potential benefit of the recognition system in the era of big data considering that real whispered emotional data is scarce. For these reasons, such strategy, i. e., deploying normal phonated speech data for whispered speech-based tasks, is adopted in this study for creating a whispered speech emotion recognition system.

Another major concern of a whispered speech emotion recognition system is: normal phonated speech fundamentally differs from whispered speech in their use of the spectrum both perceptually and for speech production. Specifically, the absence of periodic vibration of the vocal folds during production of whispered speech leads to the lack of voiced excitation, the lack of harmonic structure, and acoustic cues signaling the fundamental frequency (F0) in speech, shifted formant locations, as well as changes in formant band width (see [10], [19]–[23]). Speech emotion systems built with normal phonated speech signals are challenged, and can deliver significantly degraded performance, when they encounter whispered speech that differs from the limited conditions under which they were originally developed and 'trained'. Hence, such differences between the test data and training data make whispered speech emotion recognition a very challenging task.

Motivated by feature transfer learning, this study will show that such concept considerably benefits a emotion recognition system for whispered speech when it uses normal phonated speech data for training as well. Specifically, this work tends to result in the transformation from the normal phonated speech domain to the whispered speech domain despite its ignorance of whispered data labels. This resulting transformation can alleviate the disparity between them and then support effective supervised learning in building a whispered speech emotion recognition system. Accordingly, the focus of the present work is placed on exploring standard but powerful feature transfer learning techniques based on autoencoders including *denoising autoencoders* (DAE) [24] , its more recent variant, i. e., *shared-hidden-layer autoencoders* (SHLA) [25], and *extreme learning machine autoencoders* (ELM-AE) [26]. As a result, the proposed feature transfer learning methods successfully endow a speech emotion model that can adapt to a range of speech modalities, i. e., normal phonated speech and whispered speech.

In addition to the motivation provided above, the core contributions of this paper can be summarized as follows:

1) To the best of our knowledge, this is the first work focusing transfer learning on whispered speech emotion.
2) Technically, we propose the autoencoder-based feature transfer learning framework, allowing us to use efficient normal phonated data to reliably recognize emotions from whispered speech.
3) For the first time, acoustic feature analysis is conducted on a whispered speech database to show which features derived from different speech modes are important for the task of interest.
4) We compare our transfer learning method with other prominent methods for whispered speech emotion recognition. We subject this method to thorough evaluation two emotional databases. Extensive experimental results show our proposed method is feasible and effective on whispered speech emotion recognition.

The remainder of this paper is organized as follows. In Section III, we first present the feature transfer learning methods based on DAE, SHLA, and ELM-AE. Section IV introduces the selected corpora for evaluation. Next, we describe the empirical evaluation in Section V, including acoustic features and the experimental setup. Experimental results on the selected databases are demonstrated in Section VI before concluding this paper in Section VII.

II. RELATED WORK

There has been a considerable amount of the relate work to overcome the problem of training/test feature distribution mismatch in the field of speech emotion recognition [27], [28]. Busso *et al.* [27] proposed an iterative feature normalization scheme designed to reduce the speaker variability, while preserving the signal information critical to discriminate between emotional states. Furthermore, the work [28] analyzed how speaker variability affects the feature distribution in detail and further a speaker normalization approach based on joint factor analysis to compensate for some of the effects identified. Recently, transfer non-negative matrix factorization with the maximum mean discrepancy algorithm was proposed to address the discrepancies between the training and test data [29].

In addition, one generic approach for reducing the mismatch problem in speech emotion recognition is known as importance weights methods. Their essential idea is to assign more weight to those training examples that are most similar to the test data, and less weight to those that poorly reflect the distribution of the test data. With this idea, Kanamori et al. proposed unconstrained least-squares importance fitting (uLSIF) to estimate the importance weights by a linear model [30]. Additionally, Sugiyama et al. modeled the importance function by a linear (or kernel) model, which resulted in a convex optimization problem with a sparse solution, called the Kullback-Leibler importance estimation procedure, or KLIEP [31]. Kernel mean matching (KMM) was proposed to directly estimate the resampling weights by matching training and test distribution feature means in a reproducing kernel Hilbert space [32]. The three methods have recently been shown to lead to significant improvement in speech emotion recognition when Hassan et al. first considered to explicitly compensate for acoustic and speaker differences between training and test databases [33].

Another possible solution to address the problem of these differences is to deploy *feature learning* (or representation learning). Feature learning, i. e., learning some transformations of the data that make it easier to extract useful information when building classifiers or other predictors, has been considered from many perspectives within the realm of machine learning [25], [34]–[36]. The key idea of feature learning is to make use of deep architectures, resulting in abstract representation. Generally, more abstract concepts are invariant to most local changes of the input. Following the concept of feature learning, *feature transfer learning* has been proposed to deal with the problem of how to reuse the knowledge learned previously from 'other' data or features [37]. This rather essential characteristic suggests that feature transfer learning would be well suited for the scenarios where the data distribution in the test domain is different from the one in the training domain but the task remains the same [25], [35]. For example, Deng et al. proposed feature transfer learning based on a sparse autoencoder method for discovering knowledge in acoustic features from small labeled target data to improve performance of speech emotion recognition when applying the knowledge to source data [35]. More recently, Mao et al. proposed a transfer learning method called sharing priors between related source and target classes based on a two-layer neural network [38]. Huang et al. proposed a novel feature transfer approach with PCANet (a deep network), which extracts both the domain-shared and the domain-specific latent features to facilitate performance improvement [39].

III. METHODS

This paper devises a system of recognizing emotional states from whispered speech mainly inspired by feature learning. Those methods take an advantage of the composition of multiple non-linear transformations of the data to generate more abstract and more useful representations, which lead to compensating for the mismatch between the training (normal phonated) and test (whispered) data. It was empirically observed that feature learning often yielded better representations, e. g., in terms of classification accuracy, quality of the samples generated by a probabilistic model or in terms of the invariant properties of the learned features [26], [34]. For example, deep neural networks, a typical approach in feature leaning, has shown to outperform traditional Gaussian mixture models (GMMs) on a variety of speech recognition benchmarks, sometimes by a large margin [40].

For feature learning, it is common to employ autoencoders to learn a new transformation at the higher level from the previously learned transformation in an unsupervised way. Autoencoders are principally developed as multi-layer-perceptrons (MLPs) with only one hidden layer predicting their inputs [41]–[43]. The autoencoder structure aims to explicitly define a direct feature encoding function in a specific parametrized closed form. This function, denoted as f, allows the simple but efficient computation of a new representation from its input. Given an example $\mathbf{x}$, we define

$$\mathbf{h} = f(\mathbf{x}), \tag{1}$$

where $\mathbf{h}$ is the representation computed from the input $\mathbf{x}$.

In light of the idea of yielding abstract representations, this section introduces in detail a whispered speech emotion system equipped with unsupervised feature learning techniques based on autoencoders. To give a comprehensive and full discussion of autoencoder-based feature transfer learning, this study covers two different methods of training an autoencoder. We first introduce DAE and SHLA trained as neural networks with back propagation (BP), and further, according to extreme learning machine (ELM) theory,

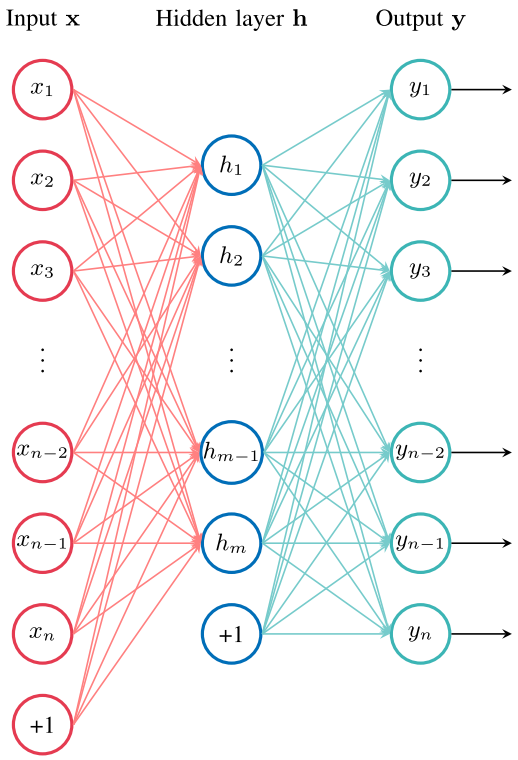

FIGURE 1. An autoencoder architecture.

present ELM-AE with a fast learning speed and good generalization capability [26]. The aim of using the three autoencoders for the creation of whispered speech emotion recognition systems is to yield robust features in undoing the effects of the mismatch between training data (normal phonated speech) and test data (whispered speech). It is then straightforward to use the resulting features as input to a 'standard' supervised classifier, such as support vector machines (SVMs).

A. AUTOENCODERS

Autoencoders, aka single-hidden layer feedforward neural networks, are illustrated in FIGURE 1. Formally, in response to an input example $\mathbf{x} \in \mathbb{R}^n$, the hidden representation $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^m$ is

$$\mathbf{h}(\mathbf{x}) = f\left(\mathbf{W}^{(1)} \cdot \mathbf{x} + \mathbf{b}^{(1)}\right), \tag{2}$$

where $f(\cdot)$ is specified as an activation function (typically a logistic sigmoid function or hyperbolic tangent non-linearity function applied component-wise), $\mathbf{W}^{(1)} \in \mathbb{R}^{m \times n}$ is a weight matrix, and $\mathbf{b}^{(1)} \in \mathbb{R}^m$ 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 $\mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{y} = g\left(\mathbf{W}^{(2)} \cdot \mathbf{h}(\mathbf{x}) + \mathbf{b}^{(2)}\right), \tag{3}$$

where $g(\cdot)$ is specified as an activation function, and $\mathbf{W}^{(2)} \in \mathbb{R}^{n \times m}$ is a weight matrix, and $\mathbf{b}^{(2)} \in \mathbb{R}^n$ is a bias vector.

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

$$\mathcal{J}_{\text{AE}}(\theta_{\text{AE}}) = \sum_{\mathbf{x} \in \mathbf{X}} \left\| \mathbf{x} - \mathbf{y} \right\|^2. \tag{4}$$

The minimization is usually realized either by BP with stochastic gradient descent or more advanced optimization techniques such as the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) or conjugate gradient method, or by ELM methods.

B. DENOISING AUTOENCODERS (DAE)

A DAE is trained to reconstruct a clean 'repaired' input from an artificially corrupted version [24]. In doing so, the learner must capture the structure of the input distribution in order to optimally reduce the effect of the corruption process, with the reconstruction essentially being a nearby but higher density point than the corrupted input [34]. Consequently, more robust features are learned in this way as compared to a basic autoencoder. Due to the useful characteristic, the DAE has been broadly considered to efficiently help speech emotion recognition [25], [36], [44].

C. SHARED-HIDDEN-LAYER AUTOENCODERS (SHLA)

The idea behind transfer learning is to exploit commonalities between different learning tasks in order to share statistical strength, and transfer knowledge across tasks [34], [45]. Based on the motivation of the 'sharing idea' in transfer learning, an alternative structure of an autoencoder that attempts to minimize the reconstruction error on both the training set and the test set was recently proposed [25]. The 'shared-hidden-layer autoencoder' (SHLA for short) shares the same parameters for the mapping from the input layer to the hidden layer, but uses independent parameters for the reconstruction process.

Given a training set of examples $\mathbf{X}_{\text{tr}}$, and a test set of examples $\mathbf{X}_{\text{te}}$, the two objective functions, analogous to the basic autoencoder's objective, are formed as follows:

$$\mathcal{J}_{\text{tr}}(\theta_{\text{tr}}) = \sum_{\mathbf{x} \in \mathbf{X}_{\text{tr}}} \left\| \mathbf{x} - \mathbf{y} \right\|^2, \tag{5}$$

$$\mathcal{J}_{\text{te}}(\theta_{\text{te}}) = \sum_{\mathbf{x} \in \mathbf{X}_{\text{te}}} \left\| \mathbf{x} - \mathbf{y} \right\|^2, \tag{6}$$

where the parameters $\theta_{\text{tr}} = \left\{\mathbf{W}^{(1)}, \mathbf{W}_{\text{tr}}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}_{\text{tr}}^{(2)}\right\}$, and $\theta_{\text{te}} = \left\{\mathbf{W}^{(1)}, \mathbf{W}_{\text{te}}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}_{\text{te}}^{(2)}\right\}$ share the same parameters $\left\{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\right\}$.

Further, we optimize the joined distance for the two sets, which leads to the following overall objective function:

$$\mathcal{J}_{\text{SHLA}}(\theta_{\text{SHLA}}) = \mathcal{J}_{\text{tr}}(\theta_{\text{tr}}) + \frac{\lambda_{\text{tr}}}{2} \left\| \mathbf{W}^{(1)} \right\|_F^2 + \left\| \mathbf{W}_{\text{tr}}^{(2)} \right\|_F^2 + \gamma \left(\mathcal{J}_{\text{te}}(\theta_{\text{te}}) + \frac{\lambda_{\text{te}}}{2} \left\| \mathbf{W}_{\text{te}}^{(2)} \right\|_F^2 \right), \tag{7}$$

where $\theta_{\text{SHLA}} = \left\{\mathbf{W}^{(1)}, \mathbf{W}_{\text{tr}}^{(2)}, \mathbf{W}_{\text{te}}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}_{\text{tr}}^{(2)}, \mathbf{b}_{\text{te}}^{(2)}\right\}$ are the parameters to be optimized during training, and the hyper-parameter γ controls the strength of the regularization. Here, $\|\mathbf{A}\|_F = \sqrt{\sum_i \sum_j |a_{ij}|^2}$ is the Frobenius norm defined as the square root of the sum of the absolute squares of a given matrix's elements. To avoid overfitting, we also include a weight-decay regularization term with its hyper-parameter λ to the objective function above.

The SHLA model can be regarded as an instance of multitask learning [46]. By explicitly adding the regularization term from the target (test) set, the SHLA is equipped with extensive flexibility to directly incorporate the knowledge of the interest task. Hence, during minimizing the objective function, the shared hidden layer is biased to make the distribution induced by the training set as similar as possible to the distribution induced by the target set. This helps to regularize the functional behavior of the autoencoder. Ultimately, it lessens the effects of the difference in the training and target sets.

D. EXTREME LEARNING MACHINES AUTOENCODERS (ELM-AE)

Recently, extreme learning machine (ELM) has been proposed since traditional BP algorithms for neural networks always converge to local optima and suffer from slow convergence. In ELM, the hidden nodes are randomly initiated and then fixed without iteratively tuning. The only trainable parameters are the weights between the hidden layer and the output layer. In this way, ELM is treated as a *linear-in-the-parameter* model which turns out to solve a linear system. The advantages of ELM in efficiency and generalization performance over traditional BP algorithms have been demonstrated on a wide range of problems from different fields [47]. Extreme learning machines autoencoders (ELM-AE) are a special case of ELM, where the input is equal to the output. In [26], Kasun *et al.* showed empirically that ELM-AE is comparable to DAE and other DNN frameworks for a handwritten digit recognition task on the MNIST data.

In contrast to DAE and SHLA, ELM-AE randomly generates the hidden nodes $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$, and only tunes the weights of the output layer $\mathbf{W}^{(2)}$ in the training phase. With (2), the input data is mapped to ELM random feature space. Given L training samples of $\mathbf{X}$, the outputs of ELM-AE turn to be as follows:

$$\mathbf{W}^{(2)}\mathbf{H} = \mathbf{X}, \tag{8}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$ are the input data, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]$ are the hidden representations. The output weights are calculated by

$$\mathbf{W}^{(2)} = \mathbf{X}\mathbf{H}^{\dagger}, \tag{9}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$.

To improve the generalization capability and further obtain the solution faster, we can add a regularization term C as suggested in [48]

$$\mathbf{W}^{(2)} = \mathbf{X}\mathbf{H}^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T\right)^{-1}, \tag{10}$$

where $\mathbf{I}$ is an identity matrix. Note that if the size of the input layer is not equal to the number of hidden units (i. e., $n \neq m$) we use (10) to compute the output weights, otherwise we use (9).

In the sense of ELM-AE, hidden nodes are important to learning but do not need to be tuned and can be independent of training data, but output weights corresponds to building the transformation from the feature space to input data. It turns out that the representations $\tilde{\mathbf{H}}$, which are learned by ELM-AE, are defined via the weights $\mathbf{W}^{(2)}$ by

$$\tilde{\mathbf{H}} = \left(\mathbf{W}^{(2)}\right)^T \mathbf{X}. \tag{11}$$

E. RECOGNITION WITH AUTOENCODER-BASED FEATURE TRANSFER LEARNING

In forming a whispered speech emotion system with normal phonated speech data, as discussed in Section I, we are faced with a scenario where the normal phonated speech data used to train a classifier has some properties that are different from the whispered speech data seen in the testing phase. Naturally, the difficulty of this system comes down to addressing this 'data bias' issue. There have been a small number of attempts at this issue. For instance, recent studies proposed various transfer learning methods to alleviate the data bias issue for speech emotion recognition [25], [33], [35], [49]. In these studies, the data bias issue is mainly caused by a change in acoustic signal conditions, or different speakers, and the type of different languages. However, the profound mismatch between different speech modes has not been considered yet. Most importantly, this work enables a speech emotion system working for whispered speech by deploying autoencoder-based feature transfer learning. Hence, this study continues to extend the autoencoder-based feature learning framework, and further, includes ELM-AE into such framework for the first time.

In detail, this work proposes three feature transfer learning methods with the integration of a DAE, an SHLA, and an ELM-AE. A central idea is to look for a transformation which would not only automatically capture useful features hidden in data, but also transfer the knowledge from the target domain (test) to the source domain (training). Algorithm 1 presents the autoencoder-based feature transfer learning methods.

The first method uses a DAE as the feature transformation to accomplish the goal of building the recognition model. In order to discover the knowledge from the test data, it is necessary to access the test data and feed it into the training procedure of a DAE (cf. Section III-A). According to the feature encoding function (cf. Section III-A), then, the optimized parameters of the DAE $\left\{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\right\}$ lead both the training

Algorithm 1 Autoencoder-Based Feature Transfer Learning

Input: The labeled training data $\mathbf{X}_{\text{tr}}$, the corresponding labels $\mathcal{L}_{\text{tr}}$, and the test data $\mathbf{X}_{\text{te}}$.

Output: Predictions $\mathcal{P}$ for the target task.

if Method 1 (*DAE*) **then**
  Train a DAE using $\mathbf{X}_{\text{te}}$ without supervision and result in $\{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\}$.
else if Method 2 (*SHLA*) **then**
  Train an SHLA using $\mathbf{X}_{\text{te}}$ and $\mathbf{X}_{\text{tr}}$ together without supervision and result in $\{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\}$.
else if Method 3 (*ELM-AE*) **then**
  Train an ELM-AE using $\mathbf{X}_{\text{te}}$ without supervision and result in $\mathbf{W}^{(2)}$.
end if
Generate new representations $\mathcal{H}_{\text{tr}}$ and $\mathcal{H}_{\text{te}}$ for $\mathbf{X}_{\text{tr}}$ and $\mathbf{X}_{\text{te}}$ (cf. Section III-A) if DAE and SHLA, otherwise Equation (11):
Learn a classifier $\mathcal{C}$ with $\mathcal{H}_{\text{tr}}$ and $\mathcal{L}_{\text{tr}}$ by applying a supervised learning algorithm (e. g., an SVM).
Make predictions $\mathcal{P}$ with $\mathcal{H}_{\text{te}}$ by the classifier $\mathcal{C}$.
return The predictions $\mathcal{P}$.

data and test data to generating their new representations. After that, these representations are taken to build a standard supervised classifier. During yielding the feature transformation, however, this method apparently ignores an attempt to explore the information behind the training data, and forces the training data to generate its new representation under the characteristics as given by the test data. In this case, we may unexpectedly lose those instances of the training data that are not following these characteristics, such that we may lose information useful for the subsequent supervised classifier to a certain degree.

By contrast, in a 'win-win strategy', we propose the second, i. e., SHLA method, to learn the common knowledge of the training data and the test data simultaneously. This method applies the training data and test data in the training of an SHLA (cf. Section III-C) so as to result in the feature transformation which would in particular balance the 'conflicts' between the two mismatched data in an optimization way. Subsequently, this method proceeds with the same steps as the first method to yield the new representations and train the classifier.

Finally, we turn to using ELM-AEs introduced in Section III-D for building a feature transfer learning framework and in turn achieve more robust features to classification. Using the test data, such method, called ELM-AE, creates the transformation with the learned output weights for compensation for the mismatch between the training domain and the test domain.

In the following, the three methods are referred to as DAE, SHLA, and ELM-AE. A great advantage of using these methods is that they take the help from the test data to optimize all parameters. This is very useful if a large amount of the test data is available in contrast to the training data. Further, abundant test data allows to apply cross-validation for tuning the hyper-parameters.

TABLE 1. Overview of the selected databases.

[#]	Emotion[a]				Valence[b]		Arousal[c]	
	A	F	H	N	−	+	−	+
GeWEC:								
Normal	160	160	160	160	320	320	160	480
Whispered	160	160	160	160	320	320	160	480
EMO-DB:								
Normal	127	55	64	78	182	142	78	246

[a] Emotion categories: anger (A), fear (F), happiness (H), and neutral (N).
[b] Binary valence: negative (−), positive (+).
[c] Binary arousal: low (−), high (+).

IV. SELECTED CORPORA

A. GENEVA WHISPERED EMOTION CORPUS

This study employs the Geneva Whispered Emotion Corpus (GeWEC) to evaluate the effectiveness of the proposed system. The corpus provides normal phonated/whispered paired utterances. Two male and two female professional French-speaking actors in Geneva were recruited to speak eight predefined French pseudo-words (*"belam"*, *"molen"*, *"namil"*, *"nodag"*, *"lagod"*, *"minad"*, and *"nolan"*) with a given emotional state in both normal phonated and whispered speech modes. Particularly, speech was expressed in four emotional states: *angry*, *fear*, *happiness*, and *neutral*. The actors were requested to express each word in all four emotional states five times. The utterances were labeled based on the state they should be expressed in, i. e., one emotion label was assigned to each utterance. As a result, GeWEC consists of 1 280 instances in total. To give an in-depth evaluation of the proposed method, we decided to further generate labels for binary valence/arousal from the emotion categories. In valence space, angry and fear have negative valence, happiness and neutral have positive valence. In arousal space, neutral is low arousal, angry, happiness and fear are high arousal. Moreover, an overview of the selected corpus is found in Table 1.

Recording was done in a sound proof chamber using professional recording equipment. All recordings were recorded with a 16 bit PCM encoded single channel at a sampling rate of 44.1 kHz. The distance from the microphone was about 0.5 m during recording. Recordings were accompanied by visual cues on a screen, which indicated which word has to be vocalized and which emotional state needs to be expressed. Cues were on the screen for 1 s length, separated by a blank screen of 2 s. The cue duration of 1 s was chosen such that the actors were guided to vocalize each word with a duration of about 1 s, which ensures that the vocalizations were comparable in length.

Pre-processing steps were applied to each utterance before feature extraction, in which all utterances were normalized

to mean energy, as well as scaled to a mean of 70 dB sound pressure level (SPL) and added manually a fade-in/fade-out duration of 15 ms.

B. BERLIN EMOTIONAL SPEECH DATABASE

A further well known set for normal phonated speech emotion classification, Berlin Emotional Speech Database (EMO-DB) [50], is chosen to test the effectiveness of the proposed methods. It covers *anger*, *boredom*, *disgust*, *fear*, *happiness*, *neutral*, and *sadness* as speaker emotions. The spoken content is again pre-defined by ten German emotionally neutral sentences like *"Der Lappen liegt auf dem Eisschrank" (The cloth is lying on the fridge.)*. Ten (five female) professional actors speak ten sentences. The actors were asked to express each sentence in all seven emotional states. The sentences were labeled according to the state they should be expressed in, i. e., one emotion label was assigned to each sentence. While the whole set comprises around 900 utterances, only 494 phrases are marked as minimum 60 % natural and minimum 80 % agreement by 20 subjects in a listening experiment. This selection is usually used in the literature reporting results on the corpus (e. g., [51]–[53]). Further we only retain those emotional states appearing in the GeWEC data for the experiments. In this way, EMO-DB in this article ends up consisting of 322 utterances as shown in Table 1.

V. EMPIRICAL EVALUATION

A. ACOUSTIC FEATURES

As for acoustic features, we chose to use a standardized feature set as is provided by the INTERSPEECH 2009 Emotion Challenge [54] which contains 12 functionals applied to 2 × 16 acoustic LLDs including their first order delta regression coefficients In detail, the 16 LLDs are MFCC 1–12, root mean square (RMS) frame energy, zero-crossing-rate (ZCR) from the time signal, probability of voicing from autocorrelation function, and pitch frequency F0 (normalized to 500 Hz). Then, 12 functionals – arithmetic mean, moments including standard deviation (SD), kurtosis and skewness, four extremes (i. e., minimum and maximum value, relative position, and ranges) as well as two linear regression coefficients with their mean square error (MSE) – are applied to the LLDs and their deltas. 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 [55], [56], which has matured to be a standard for feature extraction in speech emotion recognition, was used with the pre-defined challenge configuration in the paper. Please note that, although a variate of moments of the LLDs are used to represent the speech signal, it is trivial to use such features in an online manner. As an example, [57] has investigated the feasibility and reliability of using such acoustic features in a distributed system for multiple Computational Paralinguistics tasks.

B. EXPERIMENTAL SETUP

We use unweighted average recall (UAR) as a performance metric, which has also been the competition measure of the first challenge on emotion recognition from speech [54] and follow-up ones. It equals the sum of the recalls per class divided by the number of the classes, and appears more meaningful than overall accuracy in the given case of presence of class imbalance. As for the basic supervised learner in the classification step, we used the $L2$-regularized $L2$-loss support vector classifier implemented in LIBLINEAR [58], with a fixed penalty factor $C = 0.5$. Besides, we always chose logistic sigmoid functions as the activation function for autoencoders.

For appropriately selecting the hyper-parameters of the autoencoders, we adopt k-fold cross validation. Therefore, the training set is split into four folds ($k = 4$) and each model is trained four times with a different fold held out as validation data. The predictions made by the four models are used to obtain a UAR when we report test set results. According to the performance on the validation data, we choose the best particular model in each family of models.

For optimization of the parameters in the autoencoders such as DAE and SHLA, we applied the third party software minFunc[1] implementing L-BFGS gradient descent. In our experiments, attempted hyper-parameters for DAE and SHLA are the following: the maximum iteration number $iter_{max} \in \{20, 40, 50, 100, \ldots, 300\}$, the number of hidden units $m \in \{64, 128, \ldots, 1\,024\}$, the weight decay value $\lambda_{tr}(\lambda_{te}) \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, and the hyper-parameter for SHLA $\gamma \in \{0, 0.1, \ldots, 1\}$. In addition to them, masking noise with a variance of 0.01 is injected to inputs during the training of the DAE and the SHLA. For ELM-AE, the number of hidden units $m \in \{64, 128, \ldots, 1\,024, 2\,000, \ldots, 7\,000\}$ and the regularization term $C \in \{10^{-5}, 10^{-4}, \ldots, 10^{8}\}$ are attempted.

VI. RESULTS

A. ACOUSTIC FEATURE ANALYSIS

This section conducts feature selection as acoustic feature analysis on GeWEC to show which features derived from different speech modes are important for the task of interest. By the means of one feature selection algorithm for ranking using the information gain with respect to the class implemented in the WEKA toolkit [59], we compare the features obtained on the normal phonated speech with those obtained on the whispered speech from the GeWEC data in FIGURE 2.

For all tasks, it can be observed that the relative importance of LLDs remarkably differs between speech modes. For instance, the F0-related features are crucial for normal phonated speech while those are completely of redundancy for whispered speech, which is expected due to the absence of the fundamental frequency in whispered speech. Besides, the probability of voicing and ZCR for whispered speech become much more reliable in the emotion and arousal cases.

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

FIGURE 2. **Full INTERSPEECH 09 feature set (Full) vs. 50 best features selected by measuring the information gain with respect to the class on whispered (W) speech and normal phonated speech (N) in the GeWEC data for emotion, valence, and arousal classification. The percentage of selected low-level descriptors (LLDs) and types of functionals is shown.**

TABLE 2. **Recognition results for emotion categories/binary valence/binary arousal in leave-one-speaker-out testing for different train/test combinations.**

UAR [%]	Train on		
Test on	Normal	Whispered	Both
Normal	74.1/73.1/58.6	41.7/61.7/60.1	58.3/70.5/61.9
Whispered	44.5/57.2/62.2	46.7/56.2/57.6	50.6/59.4/59.4
Both	59.3/65.2/60.4	44.2/59.0/58.9	59.5/64.9/60.6

One possible reason causing such change is that for whispered speech, discrimination performance is mainly affected by the high-frequency region whereas for normal phonated speech, discrimination performance is mainly affected by the low-frequency region [23]. As regards the types of functionals, it can be observed that the relative importance of means and moments increases for whispered speech when compared to normal phonated speech.

Overall, the acoustic feature analysis shows that relevant features for use in speech emotion recognition model construction are different from normal phonated speech and whisper speech, and using normal phonated speech as training set to recognize emotional states from whispered speech is very challenging.

B. BENCHMARK TESTS ON GEWEC

We first run a number of experiments where the training and the test set varies in the combinations of normal phonated speech and whispered speech within the data GeWEC. These include matched and mismatched as well as multi-condition training and testing. Table 2 lists the results of all nine different training and test set combinations. Apart from emotion categories, we also evaluate the discrimination between binary valence and the discrimination between binary arousal. For practical and meaningful comparisons, the speaker-independent leave-one-speaker-out cross validation strategy is adopted to meet speaker independent criteria.

As may be expected and can be seen from Table 2, the recognition system using supra-segmental features works best when both the training and test data are entirely drawn from normal phonated speech, leading to the largest UAR of 74.1 % for the four-class emotion classification problem. Further — also as one may expect —, whispered speech (in matched condition) reaches a significantly lower UAR of 46.7 %. Using whispered speech for training seems to downgrade in particular the recognition of valence. It seems plausible that a training set drawn from whispered speech should be a better way for whispered speech emotion recognition (i. e., matched condition learning). However, there is no significant reduction in the system using normal phonated speech based on Table 2. For binary valence and binary arousal, it is even surprisingly observed that the system trained with normal phonated speech sometimes obtains slightly higher UAR than the ones when trained with whispered speech. Further, a multi-condition training is only truly beneficial for whispered speech.

C. RESULTS FOR THE PROPOSED METHODS ON GEWEC

This section reports the results obtained by the emotion recognition system using the proposed and further domain adaptation methods.

We compare a basic model without any adaptation and the three methods above, listed as follows, to evaluate our proposed approaches:

TABLE 3. Average UAR over ten trials on GeWEC: no Transfer Learning ('none'), and methods KLIEP, uLSIF, and KMM, and the proposed autoencoder-based feature transfer learning methods DAE, SHLA, and ELM-AE. Significant results (p-value < 0.05, one-sided z-test) are marked with an asterisk, judged relative to 'none'. The best UAR is highlighted in bold. Speaker-independent classification by SVM.

GeWEC	Whispered (test), Normal (train)			Normal (test), Whispered (train)		
UAR [%]	Emotion	Valence	Arousal	Emotion	Valence	Arousal
None	45.3 ± 0.0	63.0 ± 0.0	65.1 ± 0.0	52.2 ± 0.0	62.8 ± 0.0	69.0 ± 0.0
KLIEP [30]	46.1 ± 0.6	63.8 ± 0.4	60.9 ± 0.9	56.7 ± 0.6	62.6 ± 0.6	65.4 ± 1.4
uLSIF [31]	45.1 ± 0.4	63.0 ± 0.5	64.7 ± 0.5	49.2 ± 0.2	62.9 ± 0.4	67.1 ± 0.3
KMM [32]	47.8 ± 0.0	62.8 ± 0.0	65.0 ± 0.0	55.8 ± 0.0	66.6 ± 0.0	72.7 ± 0.0
DAE	$*53.7 \pm 1.6$	63.6 ± 1.1	67.2 ± 2.1	53.1 ± 1.7	66.9 ± 1.8	$*76.4 \pm 3.9$
SHLA	$\mathbf{*54.5 \pm 1.6}$	63.5 ± 1.2	$*70.6 \pm 2.9$	$*58.3 \pm 1.8$	66.0 ± 1.9	$*81.1 \pm 5.2$
ELM-AE	$*52.3 \pm 0.4$	$\mathbf{65.0 \pm 0.9}$	$\mathbf{*74.6 \pm 1.0}$	$\mathbf{*63.7 \pm 0.7}$	$\mathbf{*72.9 \pm 0.7}$	$\mathbf{*85.6 \pm 0.2}$

- 'None': employs a conventional speech emotion system, i. e., involving no adaptation, to predict emotions for a given whispered utterance.
- KLIEP [30], uLSIF [31], and KMM [32]: utilize these modern domain adaptation methods for covariate shift adaptation, respectively before SVM classification. We chose the 'tuning parameters' following [33].

First we train speech emotion recognition models on normal phonated speech while testing on whispered speech. Because of the random initialization in the autoencoders and IW methods, the results of the averaged UAR over ten trials, along with significance level computed by a one-sided z-test, are given in Table 3. We found that the DAE, SHLA, and ELM-AE outperform all the other approaches. In detail, the best performing methods for the three tasks, which achieve UARs of 54.5 %, 65.0 %, and 74.6 %, respectively, use autoencoders. For all the three tasks, the IW methods just achieve similar results as the None. On two of the three tasks, however, the autoencoder-based methods exhibit a statistically significant improvement over the None. Note that the SHLA improves on the DAE, showing that it can leverage information both from the training set and the test set in a more effective way. In the meantime, the ELM-AE generally outperforms the SHLA, which may indicate that the ELM-AE tends to attain more generalization performance.

To further test the effectiveness of the proposed methods at reducing the mismatch problem, more experiments for recognizing emotions from normal phonated speech are considered, specifically in which training data is whispered speech, and test data is normal phonated speech. Table 3 also summarizes these results. It shows that the proposed methods consistently outperform all the other methods since they achieve the highest UARs for the three tasks as well. In other words, they are also found effective for the normal phonated speech emotion recognition systems when a mismatch is given.

D. RESULTS FOR THE PROPOSED METHODS ON EMO-DB

Although the system is carefully designed for whispered speech, it would be also expected to see if such system can be suitable for normal phonated speech since normal phonated speech is a more common way in our daily life. Therefore, we further test the proposed methods on normal phonated speech. In doing so, the recognition models obtained by the proposed methods as well as other methods for comparison, which are originally developed for whispered speech in Section VI-C, continue to make predictions on the EMO-DB data. Note that these models are only trained with normal phonated speech data from the GeWEC data in order to classify whispered speech. Following the experimental settings above, we present these results in Table 4.

TABLE 4. Average UAR on EMO-DB: All the models are originally developed for whisper speech in Section VI-C. These models are only trained with normal phonated speech from the GeWEC data in order to classify whispered speech.

UAR [%]	Emotion	Valence	Arousal
None	49.9 ± 0.0	$\mathbf{84.2 \pm 0.0}$	75.0 ± 0.0
KLIEP [30]	17.1 ± 0.9	66.4 ± 0.6	34.9 ± 0.7
uLSIF [31]	19.7 ± 0.6	67.8 ± 0.3	34.1 ± 0.3
KMM [32]	15.8 ± 0.0	68.2 ± 0.0	27.4 ± 0.0
DAE	54.5 ± 1.5	72.2 ± 1.9	78.6 ± 3.9
SHLA	55.4 ± 1.6	69.5 ± 1.8	$*82.1 \pm 1.8$
ELM-AE	$\mathbf{*57.4 \pm 0.4}$	76.0 ± 0.3	$\mathbf{*85.5 \pm 1.0}$

We found that the proposed methods can retain the competing performance as the None, where the training and test data come from normal phonated speech, whereas all of the IW methods lead to a significant reduction in performance. In addition, for emotion and arousal tasks, the proposed methods significantly improve the performance in UAR over the None, which may indicate that the knowledge of whispered speech automatically found by the proposed methods might be beneficial for normal phonated speech recognition to some degree. Overall, these findings may suggest that the autoencoder-based methods have great advantages to generate feature representations which are common to or invariant across both whispered and normal phonated speech.

E. COMPARISON BETWEEN THE AUTOENCODER-BASED METHODS

FIGURE 3 demonstrates how the number of hidden units m influences the performance of the different autoencoder-based methods on GeWEC and EMO-DB. It can

TABLE 5. Comparison of running time (s) of DAE, SHLA, and ELM-AE on GeWEC.

Hidden units	64	128	256	512	1 024
DAE	4.734 ± 0.500	9.098 ± 2.808	14.023 ± 2.904	23.062 ± 4.828	35.952 ± 11.126
SHLA	11.348 ± 2.091	15.799 ± 1.543	27.159 ± 3.885	42.896 ± 8.322	61.502 ± 14.604
ELM-AE	0.020 ± 0.005	0.035 ± 0.005	0.086 ± 0.0569	0.154 ± 0.029	0.337 ± 0.040

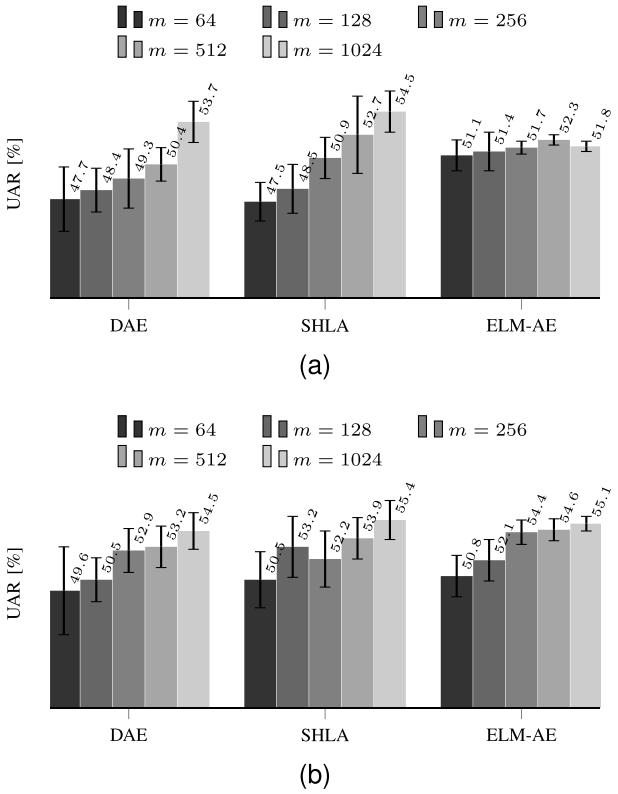

FIGURE 3. Average UAR with standard deviation over ten trials obtained by DAE, SHLA, and ELM-AE with changes in the number of hidden units (*m*) for the emotion labeling scheme. (a) On GeWEC. (b) On EMO-DB.

been seen that the change in the number of hidden units, within a particular range, has a strong influence on the proposed methods. That is, we could obtain a sustained performance growth with more hidden units.

Furthermore, we compare the running time of DAE, SHLA, and ELM-AE on GeWEC. As can be seen from Table 5, the ELM-AE has the least amount of needed running time with respect to the DAE and SHLA, simply because its training phase avoids tuning the parameters iteratively.

VII. DISCUSSIONS AND CONCLUSIONS

Autoencoder-based feature transfer learning, has been applied to speech emotion recognition primarily for cross-corpus classification of emotions [25], [35], [49], rather than whispered speech classification. We extend these work by showing how autoencoder-based feature transfer learning can be applied to create a recognition engine owing a completely trainable architecture that can adapt it to a range of speech modalities, such as normal phonated speech and whispered speech.

To reach the goal of this work, i. e., developing an emotion recognition system which is trained on normal phonated speech and can offer reliable performance also for whispered emotional speech, we proposed three feature transfer learning methods using denoising autoencoders, shared-hidden-layer autoencoders, and extreme learning machines autoencoders for whispered speech emotion recognition. Our results demonstrate that such feature transfer learning methods can significantly enhance the prediction accuracy on a range of emotion tasks and compete well with other alternative methods. The proposed methods also do not reduce system performance on normal phonated speech.

We further found that autoencoder-based feature transfer learning not only can aim to alleviate the mismatch between the training set and test set by discovering common features across multiple modes or different corpora, which has been repeatedly shown in previous work like [60], [61], but also can greatly improve the learning performance of a target task by transferring useful information in one source task to the target task in an unsupervised way. Table 4 provides a piece of evidence on the point. Note that, here, whispered speech as the source obviously offers helpful information so as to improve the target task of normal phonated speech emotion recognition. Such benefit has been constantly demonstrated in other transfer learning methods and widely applied in a variant of applications such as web-document classification [62] and WiFi-based indoor localization [63], but has never been found for autoencoder-based feature transfer learning before. Hence, this work provides a new insight into the way we explore autoencoder-based feature transfer learning.

Overall, our results are very informative and encouraging for future exploitation of the whispered speech recognition system proposed in this paper. However, this work is only a first step towards the creation of whispered speech emotion recognition, and many more experiments need to be carried out in the future. Since this study using acted data for evaluation tends to lead to overestimated performance [64], we hope to perform further work on spontaneous data, which will make recognition systems even more applicable in real-life settings. Obviously, however, collecting sponteanous *whispered* emotion in large quantities will remain quite a challenge. Given the rapid progress of research on deep learning, we also hope to apply deep structures to the whispered speech

emotion recognition problem and the transfer learning we proposed.

REFERENCES

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, 2011.

[2] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[4] 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, no. 2, pp. 155–177, 2012.

[5] F. H. Knower, "Analysis of some experimental variations of simulated vocal expressions of the emotions," *J. Social Psychol.*, vol. 14, no. 2, pp. 369–372, 1941.

[6] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[7] P. Mitev and S. Hadjitodorov, "Fundamental frequency estimation of voice of patients with laryngeal disorders," *Inf. Sci.*, vol. 156, nos. 1–2, pp. 3–19, 2003.

[8] Y. Jin, Y. Zhao, C. Huang, and L. Zhao, "Study on the emotion recognition of whispered speech," in *Proc. GCIS*, Xiamen, China, 2009, vol. 3, pp. 242–246.

[9] G. Chenghui, Z. Heming, Z. Wei, W. Yanlei, and W. Min, "A preliminary study on emotions of Chinese whispered speech," in *Proc. IFCSTA*, Chongqing, China, 2009, vol. 2, pp. 429–433.

[10] X. Fan and J. H. L. Hansen, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Commun.*, vol. 55, no. 1, pp. 119–134, 2013.

[11] C. Zhang and J. H. L. Hansen, "An advanced entropy-based feature with a frame-level vocal effort likelihood space modeling for distant whisper-island detection," *Speech Commun.*, vol. 66, pp. 107–117, Feb. 2015.

[12] J. Zhou, R. Liang, L. Zhao, L. Tao, and C. Zou, "Unsupervised learning of phonemes of whispered speech in a noisy environment based on convolutive non-negative matrix factorization," *Inf. Sci.*, vol. 257, pp. 115–126, Feb. 2014.

[13] T. Irino, Y. Aoki, H. Kawahara, and R. D. Patterson, "Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination," *Speech Commun.*, vol. 54, no. 9, pp. 998–1013, 2012.

[14] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 2598–2602.

[15] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech," *J. Voice*, vol. 26, no. 2, pp. e49–e56, 2012.

[16] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Dialogues With Social Robots*. Saariselkä, Finland: Springer, Jan. 2016.

[17] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[18] S. E. Bou-Ghazale and J. H. L. Hansen, "HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 201–216, May 1998.

[19] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun.*, vol. 45, no. 2, pp. 139–152, 2005.

[20] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, nos. 7–8, pp. 515–520, 2002.

[21] M. Matsuda and H. Kasuya, "Acoustic nature of the whisper," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 133–136.

[22] S. T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acustica United Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[23] W. F. L. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French," *J. Acoust. Soc. Amer.*, vol. 135, no. 4, pp. 2026–2040, 2014.

[24] 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.

[25] 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. ICASSP*, Florence, Italy, 2014, pp. 4851–4855.

[26] L. L. C. Kasun, H. Zhou, and G.-B. Huang, "Representational learning with ELMs for big data," *IEEE Intell. Syst.*, vol. 28, no. 6, pp. 31–34, Nov. 2013.

[27] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Proc. ICASSP*, 2011, pp. 5692–5695.

[28] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models—Analysis and normalisation," in *Proc. ICASSP*, Vancouver, BC, Canada, 2013, pp. 7522–7526.

[29] P. Song *et al.*, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Commun.*, vol. 83, pp. 34–41, Oct. 2016.

[30] 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.

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

[32] 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.

[33] 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.

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

[35] 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.

[36] R. Xia and Y. Liu, "Using denoising autoencoder for emotion recognition," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 2886–2889.

[37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[38] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 2608–2612.

[39] 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.

[40] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[41] Y. LeCun, "Modeles connexionnistes de l'apprentissage (connectionist learning models)," Ph.D. dissertation, Univ. Pierre Marie Curie, Paris, France, Jun. 1987.

[42] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biol. Cybern.*, vol. 59, nos. 4–5, pp. 291–294, 1988.

[43] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proc. NIPS*, Denver, CO, USA, 1993, pp. 3–10.

[44] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 990–994.

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

[46] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[47] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Netw.*, vol. 61, pp. 32–48, Jan. 2015.

[48] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[49] 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.
[50] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1517–1520.
[51] H. Meng, J. Pittermann, A. Pittermann, and W. Minker, "Combined speech-emotion recognition for spoken human-computer interfaces," in *Proc. SPCOM*, Dubai, United Arab Emirates, 2007, pp. 1179–1182.
[52] V. Slavova, W. Verhelst, and H. Sahli, "A cognitive science reasoning in recognition of emotions in audio-visual speech," *Int. J. Inf. Technol. Knowl.*, vol. 2, no. 4, pp. 324–334, 2008.
[53] B. Schuller, M. Wimmer, L. Mosenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *Proc. ICASSP*, Las Vegas, NV, USA, 2008, pp. 4501–4504.
[54] B. W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.
[55] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM-MM*, Florence, Italy, 2010, pp. 1459–1462.
[56] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM-MM*, Barcelona, Spain, 2013, pp. 835–838.
[57] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 406–417, Oct./Dec. 2014.
[58] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.
[59] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[60] 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.
[61] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized stacked denoising autoencoders," in *Proc. ICML*, Edinburgh, Scotland, 2012, pp. 1–2.
[62] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. ICML*, Corvallis, OR, USA, 2007, pp. 193–200.
[63] S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu, "Transfer learning for WiFi-based indoor localization," in *Proc. AAAI*, Chicago, IL, USA, 2008, pp. 43–48.
[64] B. Schuller *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul./Dec. 2010.

JUN DENG received the bachelor's degree in electronic and information engineering from Harbin Engineering University, China, in 2009, the master's degree in information and communication engineering from Harbin Institute of Technology, China, in 2011, and the Ph.D. degree in electrical engineering and information technology from Technical University of Munich, Munich, Germany, in 2016, with a focus on feature transfer learning for speech emotion recognition. He is currently a Post-Doctoral Researcher with the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany. His interests are machine learning methods, such as transfer learning and deep learning, with an application preference to affective computing.

SASCHA FRÜHHOLZ received degrees in science education in 2001 and in psychology in 2006, and the Ph.D. degree from Bremen University, Germany, in 2008, with a focus on the neural mechanisms of facial expressions. He is currently a Swiss National Science Foundation Professor with the Department of Psychology, University of Zurich, Zürich, Switzerland. He is also with the Neuroscience Center Zurich, ZNZ, University of Zurich, and with ETH Zürich, and with the Zurich Center for Integrative Human Physiology, University of Zurich. His current projects deal with dynamic connectivity patterns of local and remote brain regions during affective voice processing using high-resolution brain scans, and specific connectivity modeling approaches for functional imaging data.

ZIXING ZHANG (M'15) received the master's degree in physical electronics from Beijing University of Posts and Telecommunications, China, in 2010, and the Ph.D. degree in engineering from the Institute for Human-Machine Communication, Technische Universität München, Germany, 2015. He is currently a Post-Doctoral Researcher with the University of Passau, Germany. He has authored more than 30 publications in peer-reviewed journals and conference proceedings. His research interests mainly lie in deep learning, semisupervised learning, active learning, and multitask learning, in the application of computational paralinguistics, including emotion and robust automatic speech recognition.

BJÖRN SCHULLER (SM'15) received the Diploma degree in 1999, the Ph.D. degree in automatic speech and emotion recognition in 2006, and the Habilitation and Adjunct Teaching Professorship in the subject of signal processing and machine intelligence in 2012, all in electrical engineering and information technology from Technical University of Munich. He is currently a tenured Full Professor heading the Chair of Complex and Intelligent Systems with the University of Passau, Germany, and also a Reader in machine learning with the Department of Computing, Imperial College London, London, U.K. He has authored or co-authored five books and more than 550 publications in peer-reviewed books, journals, and conference proceedings, leading to more than 11 000 citations (h-index=51). He is the President -Emeritus of the Association for the Advancement of Affective Computing, an Elected Member of the IEEE Speech and Language Processing Technical Committee, and a member of the ACM and the ISCA.