Interspeech 2018
2-6 September 2018, Hyderabad, India

Transfer Learning for Improving Speech Emotion Classification Accuracy

Siddique Latif[1,3], *Rajib Rana*[2], *Shahzad Younis*[3], *Junaid Qadir*[1], *Julien Epps*[4]

[1]Information Technology University (ITU)-Punjab, Pakistan
[2]University of Southern Queensland, Australia
[3]National University of Sciences and Technology (NUST), Pakistan
[4]The University of New South Wales, Sydney, Australia

siddique.latif@itu.edu.pk, rajib.rana@usq.edu.au, muhammad.shahzad@seecs.edu.pk, junaid.qadir@itu.edu.pk, j.epps@unsw.edu.au

Abstract

The majority of existing speech emotion recognition research focuses on automatic emotion detection using training and testing data from same corpus collected under the same conditions. The performance of such systems has been shown to drop significantly in cross-corpus and cross-language scenarios. To address the problem, this paper exploits a transfer learning technique to improve the performance of speech emotion recognition systems that is novel in cross-language and cross-corpus scenarios. Evaluations on five different corpora in three different languages show that Deep Belief Networks (DBNs) offer better accuracy than previous approaches on cross-corpus emotion recognition, relative to a Sparse Autoencoder and Support Vector Machine (SVM) baseline system. Results also suggest that using a large number of languages for training and using a small fraction of the target data in training can significantly boost accuracy compared with baseline also for the corpus with limited training examples.

Index Terms: cross-corpus, speech, emotion recognition, Deep Belief Networks

1. Introduction

In recent years, speech emotion recognition has received increasing interest. Automatic Speech emotion recognition focuses on using linguistic and acoustic attributes as input features and machine learning models as classifiers to classify the emotions of the speaker [1]. These systems achieve promising results when training and testing are performed from the same corpus [2, 3]. However, for real applications, such systems have been demonstrated not to perform well when speech utterances from different languages and different age groups, in quite different conditions, are combined [4].

At present, various emotional corpora exist, but they are dissimilar in terms of the spoken language, type of emotion (i.e., naturalistic, elicited, or acted) and labelling scheme (i.e., dimensional or categorical) [5]. There are more than 5,000 spoken languages around the world, but only 389 languages account for 94% of the world's population[1]. Even for 389 languages, very few adequate resources (speech corpus) are available for language and speech processing research. This means that research in language and speech analysis must confront the problem of data scarcity for many languages. This imbalance, variation, diversity, and dynamics in speech and language databases means that it is almost impossible to learn a model from a single corpus and then expect it to be effective in practice in general.

[1]https://www.ethnologue.com/statistics

In automatic speech emotion recognition, most studies focus on a single corpus at a time, without considering the performance of model in cross-language and cross-corpus scenarios. However, ever since transfer learning has been applied to cross-domain classification and pattern recognition problems, interest in applying it to cross-corpus emotion recognition has been growing. Transfer learning focuses on adapting knowledge from available auxiliary resources to transfer this learning to a target domain, where a very few or even no labelled data is available [6, 7].

Deep neural network (DNN) based transfer learning has recently improved image classification by using a very large dataset as source domain and small data as a target domain [8]. Inspired by this success, deep learning based transfer learning has recently been used for speech analysis. However, the existing research has focused on basic DNNs. The impact of using models like Deep Belief Networks (DBNs), which have strong generalisation power and are therefore suitable for cross-corpus emotion recognition, has not been thoroughly explored. A few studies have explored DBNs for speech emotion recognition (e.g., [9, 10]) and numerous studies focus on DBNs for features extraction [11–13] from speech signal. However, transfer learning using DBNs is very rare. Furthermore, how to maximise the transfer learning performance for cross-corpus/cross-language emotion recognition still needs to be explored further.

In this study, we address the above challenges. We investigate DBNs for transfer learning over five widely-used emotional speech databases. By using the experimental results from various scenarios, we indicated how a large gain in accuracy comparable to baseline can be achieved using transfer learning technique for cross-corpus emotion recognition.

2. Related Work

Although cross-language and cross-corpus speech emotion recognition is an interesting problem, relatively few studies have addressed this topic. Existing studies have mostly studied the preliminary feasibility of cross-corpus learning and pointed to the need for further in-depth research. For example, Schuller et al. [5] used six different corpora to analyse cross-corpora emotion recognition using Support Vector Machine (SVM) and highlighted the limitations of current systems for cross-corpus emotion recognition. Eyben et al. [14] used four corpora to evaluate some pilot experiments on cross-corpus emotion recognition while using SVM. They used three datasets for training and a fourth for testing, and showed that the cross-corpus emotion recognition is feasible. To explore the universal cues of emotions across languages, Xia et al. [15] investigated cross-language emotion recognition for Mandarin vs. Western lan-

10.21437/Interspeech.2018-1625

guages (i.e., German, and Danish). The authors focused on gender-specific speech emotion recognition and achieved the classification rates higher than the chance level but less than baseline accuracy. Albornoz et al. [16] developed an ensemble SVM for emotion detection with a focus on emotion recognition in unseen languages.

Deep learning techniques have been widely used for transfer learning in speech recognition but only basic DNN models have been utilised so far. Lim et al. [17] proposed cross-acoustic transfer learning framework by using DNNs. The authors trained a model on a large data of speech and use it for sound event classification. After a series of experiments, the results showed that the cross-acoustic transfer learning can significantly enhance the sound event classification rate. In [18], authors used a single DNN for speaker and language recognition with a large gain on performance by training the model on speech recognition data. These studies exploited the models that have good learning abilities so that the learned features are transferable to enable model adaptation regarding the target domain.

In this paper, we use Deep Belief Networks (DBNs) for transfer learning speech emotion. The key reason for employing DBN is its power of generalisation, which is not present in other DNN models [19]. Because, the building block of DBNs (i.e., RBMs) are universal approximators, and they are very powerful to approximate any distribution [20]. Intuitively, for cross-corpus and cross-language emotion recognition, the generalisation power of a model is crucial. In addition, DBN can learn more powerful and effective discriminative long-range of features [21] that have been shown to help in speech-related problems [22].

Apart from DNNs, researchers have also used interesting deep architectures for transfer learning. In [23], the authors focused on using Progressive Neural Networks to transfer knowledge for three paralinguistic tasks, i.e., emotion, speaker, and gender detection. Progressive Networks are useful for conducting multitasking in a network, however, we focus on a single task of emotion recognition as speaker and gender recognition are not the focus of this paper. Zong et al. [24] proposed a domain-adaptive least-squares regression (DaLSR) model for cross-corpus speech emotion recognition. They used three datasets for the evaluations and found that DaLSR can achieved better results than other models like SVM. They did not focus on achieving results higher than the baseline accuracy. Similarly, Deng et al. [25] used sparse autoencoders for feature transfer learning to improve the performance of speech emotion recognition. They used six standard databases and trained a single-layer sparse autoencoder for discovering knowledge from the target domain, and then apply these discovered representations to the source domain for reconstruction of class-specific data. Experiments using reconstructed data for classification improved the performance of the model for emotion recognition task.

3. Experimental Setup

3.1. Speech Databases

To investigate the performance of DBN for cross-corpora and cross-language emotion recognition, we selected five publicly available and highly popular corpora which have maximum diversity in languages. These databases are annotated differently, therefore, one of the only consistent ways to investigate transfer learning is by considering the binary positive/negative valence classification problem. We adopt the binary valence mapping per emotion category from [5,25,31]. The names of the datasets used in our experiments and their categorical mappings to binary valence classes are provided in Table 1. These databases were chosen to span a variety of languages.

3.2. Speech Features

In this study, we use eGeMAPS feature set, which is a widely used reference feature set for speech emotion recognition studies [23]. The feature set includes Low-Level Descriptor (LLD) features of the speech signal which are described most relevant to emotions by Paralinguistic studies [31]. The eGeMAPS feature set contains 88 features including frequency, energy, spectral, cepstral, and dynamic information. The overall components are the arithmetic mean and coefficient of variation of 18 LLDs, 6 temporal features, 4 statistics over the unvoiced segments, 8 functionals applied to loudness and pitch, and 26 additional dynamic and cepstral components.

3.3. Deep Belief Networks

DBNs are very popular deep architectures that consist of the stack of Restricted Boltzmann Machines (RBMs) to make a powerful probabilistic generative model by using layer-wise training in a greedy manner. RBM is an undirected stochastic neural network consisting of a visible layer, a hidden layer, and a bias unit. Each visible unit of the visible layer is fully connected to hidden units in the hidden layer, and the bias is connected to all the visible units and the hidden units. There is no connection between visible to visible and between hidden to hidden units. RBMs can also be used as classifiers. They are trained on the joint distribution of input data and corresponding labels, then the label is assigned to the new input which has the highest probability under the model. The joint distribution between visible layer (v) and hidden layer (h) is given by [32]:

$$P(v,h) = \frac{1}{Z}\exp(-E(v,h)) \tag{1}$$

where Z represents the normalisation constant and $E(v,h)$ is an energy function which is defined as:

$$E(v,h) = -\sum_{i=1}^{D}\sum_{j=1}^{k} W_{ij}v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{k} a_j h_j \tag{2}$$

where v_i and h_i are the binary states of visible and hidden units. W_{ij} represents the weights of connections between hidden and visible nodes. b_i and a_j are the bias terms for visible and hidden units respectively. The conditional probabilities for the visible and hidden units are given by the following equations:

$$P(v_i = 1|h) = g\big(b_i^v + \sum_j h_j W_{ij}\big) \tag{3}$$

$$P(h_j = 1|v) = g\big(b_j^h + \sum_i v_i W_{ij}\big) \tag{4}$$

where g is the sigmoid function:

$$g(x) = \frac{1}{1+e^{-x}} \tag{5}$$

An RBM is pre-trained for the maximisation of data log-likelihood $logP(v)$. The stack of generatively pre-trained RBMs constitutes a powerful DBN that can be discriminatively fine-tuned to improve performance. Weight initialisation with

Table 1: *Corpora information and the mapping of class labels onto Negative/Positive valence.*

Corpus	Language	Age	Utterances	Negative Valance	Positive Valance	References
FAU-AIBO	German	Children	18216	Angry, Touchy, Emphatic, Reprimanding	Motherese, Joyful, Neutral, Rest	[26]
IEMOCAP	English	Adults	5531	Angry, Sadness	Neutral, Happy, Excited	[27]
EMO-DB	German	Adults	494	Anger, Sadness, Fear, Disgust, Boredom	Neutral, Happiness	[28]
SAVEE	English	Adults	480	Anger, Sadness, Fear, Disgust	Neutral, Happiness, Surprise	[29]
EMOVO	Italian	Adults	588	Anger, Sadness, Fear, Disgust	Neutral, Joy, Surprise	[30]

pre-training can help the network to avoid poor local minima and give better discriminative results when compared with a neural network initialised by small random weights [33]. In this work, we also use layer-by-layer pre-training for DBN. The description of DBNs and their training methodologies can be reviewed in [32, 34].

During experimental work, a DBN with three RBM layers was selected, where the first two RBMs have 1000 hidden unit each, and the third RBM have 2000 hidden units with learning rate of 10^{-3} and 500 epochs. This configuration was obtained using cross validation experiments on validation data. The other network parameters were chosen by following the setup in [10, 35].

4. Results

In this section, we explore various scenarios for cross-corpus and cross-language speech emotion recognition and conduct experiments to test the scenarios.

4.1. Within Corpus Scheme

In order to obtain the baseline comparison results, we compare the performance of DBN with a popular approach of using sparse autoencoder with SVM for feature transfer learning in speech emotion recognition [25]. This preliminary experiment enables us to set maximum achievable baseline accuracy when both systems are trained and tested using the data of same corpus. For baseline experiments, 75% of randomly selected data is used for training and remaining 25% unseen data is used for testing. Figure 1 shows the comparison results, where DBN outperforms sparse AE for all databases.

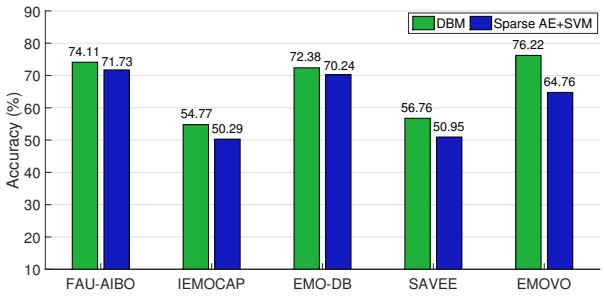
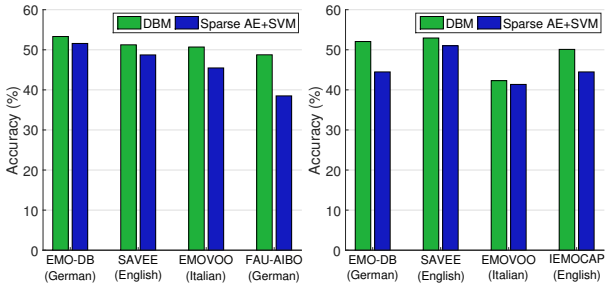


Figure 1: *Comparison of baseline accuracy using DBN and sparse AE on different databases.*

4.2. Language Tests

In this experiment, we use one language dataset for training and the remaining datasets for testing. For brevity, we just use FAU-AIBO (German) and IEMOCAP (English) datasets for training. In order to evaluate the model on IEMOCAP, we used two sessions out of five with two-fold cross validation because overall data is large. The other databases are small compared to IEMOCAP, therefore, we used them completely. Figure 2 shows the recognition rate achieved in these experiments and its comparison with previous techniques using sparse autoencoder and SVM (sparse AE+SVM) for cross-corpus transfer learning. When the IEMOCAP database was used for training the DBN, we performed pairwise testing using OHM and MONT separately for FAU-AIBO. Note that OHM and MONT are the schools whose children have participated in data formation. It can be noted from Figure 2 that DBN outperforms sparse AE for all scenarios. Beyond this point, the accuracy of sparse AE is not given, as we observe that DBNs consistently outperform sparse AE.

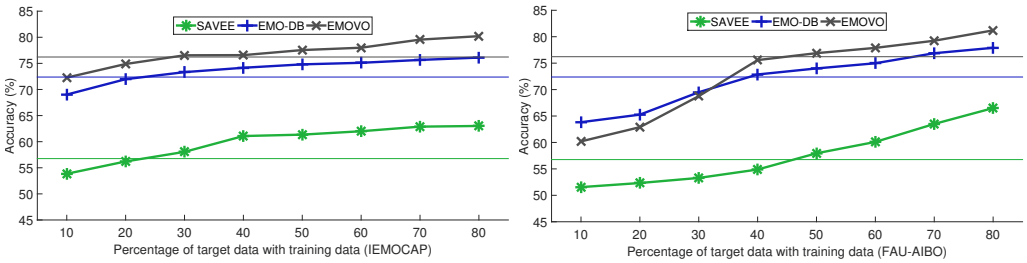


(a) (b)

Figure 2: *Comparison of language tests using DBN and sparse AE. Figure 2a represents the recognition rate using IEMOCAP (English) for training and other databases for testing whereas 2b shows the recognition rate using FAU-AIBO (German) for training and other databases for testing.*

4.3. Percentage of Target Data

In this experiment, we vary the percentage (10% to 80%) of the target dataset for the training of the model. The training was performed using IEMOCAP and FAU-AIBO separately and EMOVO, EMO-DB and SAVEE were used for testing. The results are shown in Figure 3. The straight horizontal lines in the figure show the baseline recognition rate for the respective corpora. These results show that the recognition rate significantly improves (than baseline) by including target domain data with the training data.



(a) (b)

Figure 3: *Impact of using a percentage of target date with training data. Where 3a shows the training with IEMOCAP and 3b is when training is performed using FAU-AIBO.*

4.4. Multi-language Training

In this experiment, we use multiple languages jointly for training to observe whether this improves the performance of using languages individually for training. We use both FAU-AIBO and IEMOCAP for training and remaining for testing. We also evaluate the model within the corpora. For IEMOCAP, we used three sessions (plus FAU-AIBO) for training and testing was performed using the remaining two sessions with two-fold cross validation. Similarly, for FAU-AIBO, a two-fold cross-validation was used, i.e., training on OHM (plus IEMOCAP) and evaluating on MONT and the inverse.

Further, we also performed training using a leave-one-data-out scheme. For FAU-AIBO, we have performed evaluation by using OHM and MONT independently taking the average results. In the case of IEMOCAP, we used two sessions (with two-fold cross validation) to evaluate the model. This performs better than baseline and two-language training as shown in Figure 4.

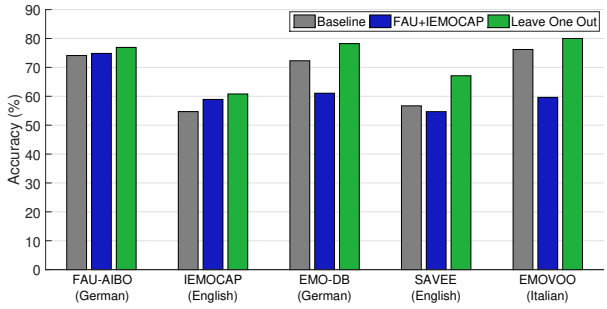


Figure 4: *Comparison of baseline results and transfer learning using FAU-AIBO+IEMOCAP and Leave-One-Out scheme.*

5. Discussion

From the experiments, Leave-One-Out seems to be standing out in-terms of obtaining the highest accuracy. This essentially means that training the model using a large range of languages would help learn many intrinsic features from each languages, which can essentially help to achieve high accuracy in an unknown language - even higher than when the same language is used for training and testing (baseline). The performance of the Leave-One-out (see Figure 4) on EMOVO database is a prime example of this. Both German and English languages have two datasets each, i.e., in a Leave-One-Out scheme there will be at least one of these language in the training set. But for EMOVO there will be a situation that emotions in the Italian language are predicted simply based on emotions in German and English language.

Another interesting aspect we learned from the experiments that including a fraction of the target data into training can help improve the performance and help achieve better results than baseline. Based on our experiments, augmenting other databases with around 20% of data (around 90 utterances in case of EMO-DB) from the target database can help achieve better than the baseline accuracy. However, this is worse while using FAU-AIBO for training. Interestingly, IEMOCAP performs well on EMO-DB that is in the German language as compared to FAU-AIBO that is also in German. We note that FAU-AIBO consists of children speech whereas EMO-DB database contains adult speech.

The performance of DBN in the language test results in Figure 2 using both IEMOCAP (English) and FAU-AIBO (German) on target datasets is poor than the baseline. The drop in accuracy is not only for the target dataset with a different language but also for target data having similar language. From this experiment, we learned that the different studio conditions, age and language differences, and type of emotional corpus cause drop in the performance of the model. This problem can be addressed by previous two findings, i.e., either by training the model with the uttrances of multiple languages or by including a small portion of data target domain with training data.

6. Conclusions

In this paper, we investigated the performance of DBNs for transfer learning based cross-corpus and cross-language speech emotion recognition. In order to evaluate the feature transference across different corpora, we performed comprehensive experiments and found that DBNs outperformed sparse autoencoders due to its increased feature learning abilities. Also, DBNs can learn from many training languages and improve the baseline accuracy even also when a small fraction of target data is included in the model while training it with a single corpus. For practical applications, these findings would be very helpful to build a robust speech emotion recognition system using data from multiple languages. Also, this would be equally useful for emotion recognition in languages with very limited or no datasets.

7. Acknowledgements

This research is partly supported by Advance Queensland Research Fellowship, reference AQRF05616-17RD2.

8. References

[1] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in *Emotion-Oriented Systems*. Springer, 2011, pp. 71–99.

[2] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[3] W. Zheng, J. Yu, and Y. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 827–831.

[4] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human emotion," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 313–323, 2012.

[5] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: a survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, 2015.

[8] Y. Sawada and K. Kozuka, "Transfer learning method using multi-prediction deep boltzmann machines for a small scale dataset," in *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*. IEEE, 2015, pp. 110–113.

[9] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 216–221.

[10] R. Rana, "Emotion classification from noisy speech-a deep learning approach," *arXiv preprint arXiv:1603.05901*, 2016.

[11] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[12] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 65–68.

[13] C. Huang, W. Gong, W. Fu, and D. Feng, "A research of speech emotion recognition based on deep belief network and svm," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[14] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions-some pilot experiments," in *Proc. LREC workshop on Emotion Corpora, Valletta, Malta*, 2010, pp. 77–82.

[15] Z. Xiao, D. Wu, X. Zhang, and Z. Tao, "Speech emotion recognition cross language families: Mandarin vs. western languages," in *Progress in Informatics and Computing (PIC), 2016 International Conference on*. IEEE, 2016, pp. 253–257.

[16] E. M. Albornoz and D. H. Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 43–53, 2017.

[17] H. Lim, M. J. Kim, and H. Kim, "Cross-acoustic transfer learning for sound event classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2504–2508.

[18] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[19] H. Lee, *Unsupervised feature learning via sparse hierarchical representations*. Stanford University, 2010.

[20] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[23] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

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

[25] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 511–516.

[26] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[29] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion(savee) database," *University of Surrey: Guildford, UK*, 2014.

[30] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database." in *LREC*, 2014, pp. 3501–3504.

[31] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[32] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[33] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010.

[34] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[35] M. A. Keyvanrad and M. M. Homayounpour, "A brief survey on deep belief networks and introducing a new object oriented toolbox (deebnet)," *arXiv preprint arXiv:1408.3264*, 2014.