The effect of personality trait, age, and gender on the performance of automatic speech valence recognition

Hesam Sagha*†, Jun Deng† and Björn Schuller*†‡
*Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany
†audEERING GmbH, Gilching, Germnay
‡Department of Computing, Imperial College London, London, UK

***Abstract*—Individual differences have significant effects on the expression of emotions. One may express the emotions openly such that they are easily recognizable, and one may be less expressive. Consequently, an emotion recognizer system will be affected by the emotion expressions from different individuals. Knowing which human factors improve or deteriorate the performance of the emotion recognizer, we can train systems based on those factors and select one of those systems that corresponds to the detected human factor of the target person. In this paper, we investigate the effect of age, gender, and Big-Five personality traits (Openness to Experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) on the performance of a speech emotion recognizer. We found that, age is the paramount factor followed by gender. Conscientiousness and Neuroticism also have a substantial effect. These findings are in congruent with the literature, meaning that the performance of a speech emotion recognizer is closely correlated with the emotion expressivity of the individuals whose speech are used for training the recognition models. Additionally, based on these findings, we create a set of simple rules to select an appropriate trained model for new speech samples. This model selection approach yields higher emotion recognition accuracy.**

1. Introduction

Emotions are being an undeniable source of information for excelling the communication between human and machines. Knowing the emotional states of speakers will help the machines to recommend better [1], adapt a system to the emotional state [2], adjust machine response dialogue [3], and so on. Therefore, having a performant emotion recognizer is an essential step toward Human-Computer Interaction.

Three sources of information have been widely studied to recognize human emotions: speech signals, facial gestures, and physiological signals. Most of the related studies conceive the produced signals –from muscular or physiological activities– as the source of the emotion expression (or displayed emotion). This point of view de-emphasizes the importance of human factors such as personality, age, and gender on the emotion expression. Scherer's appraisal model of emotions defines emotions as a consequence of components which are activated from the point an emotional stimulus is presented up to the point emotions are expressed [4]. This path engages different nervous systems and circuits in the brain, which are formed through life-experiences and practices and build different personality traits.

Recent studies reveal some relationships between induced emotions, personality, and physiological changes [5]. However, induced emotions are different from displayed emotions [6]: An emotion can be induced to a person (e. g., by showing affective pictures or music) and the person feels the emotion, but it could be without unintended particular changes in the facial muscles or vocal tract. This could be due to the individual differences or reduced affect display (e. g., as a side-effect of autism, schizophrenia, or depression). On the other hand, intended displayed emotions (such as actors' performance) are also subjective, and different individuals may or may not express emotions as they are expected. We can measure to what extent human factors affect displayed emotions by measuring the discriminability of their produced emotions. In other words, if certain personal characteristic has a positive impact on producing distinct emotions, then the emotions can be classified more accurate. For example, females (intentionally or unintentionally) can express better emotional states than males [7], [8]. Therefore, we expect higher recognition performance when an emotion recognition system is trained and tested on female subjects. By discovering the relationships between human factors and displayed emotions, it would be possible to select a specific emotion recognizer system trained on specific speaker's trait, age, or gender so as to increase the recognition performance.

In this paper, we investigate how the personality trait, age and gender could impact the performance of *speech* emotion recognizer. Additionally, by discovering these effects, we set some basic rules to select appropriate model to improve this performance.

2. Literature review

Two elements are of importance to have an accurate emotion recognition system: (i) to what extent a speaker can display his/her emotions via vocal tract or facial gestures, and (ii) how accurate is the recognition system *per se*. In this section, we review the studies on the mediation

of some individual factors on emotion expression (section 2.1) and how these factors can affect automatic emotion recognition performance (section 2.2). Apart from gender and age, we investigate Big-Five personality traits. These traits are: **O**penness to experience (Artistic, Imaginative, Insightful, Wide interests), **C**onscientiousness (Efficient, Organized, Reliable, Responsible), **E**xtroversion (vs. Introversion, Energetic, Outgoing, Talkative), **A**ggreeableness (Appreciative, Generous, Kind, Trusting), and **N**euroticism (Anxious, Tense, Touchy, Unstable).

2.1. Human factors and emotion expression

It has been shown that, when emotions are both expressed and recognized by members of the same national, ethnic, or regional group, the emotion recognition accuracy is higher [9]. Gross and John have used the self-administrating Berkeley Expressivity Questionnaire (BEQ) to measure emotion expressivity within three subscales: Negative Expressivity, Positive Expressivity, and Impulse Strength [10]. They found that Asian-Americans are less expressive than other ethnic groups, and Positive mood, Extroversion, and Agreeableness are most strongly related to the Positive Expressivity subscale [7]. Furthermore, negative mood, Neuroticism, and somatic complaints are most strongly related to the Negative Expressivity subscales. These findings are also proved by [11] where they found a significant overall positive relationship between Extroversion and emotional expressiveness and a significant overall negative relationship between Neuroticism and behavioral measures of emotional expressiveness. Furthermore, Abe and Izard deduced that full-face negative expression of a baby is directly related to Neuroticism and inversely related to Agreeableness and Conscientiousness [12]. By contrast, full-face positive expression of a baby is positively correlated with Extroversion and Openness to Experience.

Moreover, it is demonstrated that, as expected, women are more expressive than men [7], [8]. Additionally, Gross et al. found that older people express less emotional expressivity [13].

2.2. Human factors and automatic emotion recognition

To the best of our knowledge, there is no study which explicitly investigate the role of human factors on the performance of emotion recognition system. The closest study is the investigation of different languages and language-families (as part of cultural assets) on the performance of automatic emotion recognizer [14], [15], [16]. The major obstacle for investigating human factors on the automatic emotion recognition is the lack of annotated data in which both emotions and human factors are labeled. In the following, to deal with this obstacle, we apply a distribution matching method to use the information of other annotated databases and label automatically utterances in emotion corpora with different individual factors.

3. Method

To observe the effects of different human factors (personality, age, and gender) on the speech emotion recognition, we split an *emotion corpus*, into two exclusive categories: male-female, young-adult, or high-low scored personality traits. The two categories should have the same number of instances to avoid bias and class imbalance effects on the analyses. We perform these splits with the help of classifiers which are trained on other corpora (hereafter: *splitter corpora*) where the (personality, age, or gender) labels are provided. However, the recording conditions for the splitter corpora could be different from the emotion corpora. Therefore, there is a need to match the distribution of the splitter corpora with the emotion corpora. In the following section, we describe briefly the extracted features from speech signals. Then, in section 3.2 we explain how to match the distribution of the splitter and emotion corpora followed by the analyses description in section 3.3. Finally, we describe the splitter and emotion corpora we used for this study in section 3.4.

3.1. Feature extraction

We extracted 384 acoustic features as in the Interspeech 2009 Emotion Challenge using openSMILE [17]. The features consist of 12 functionals of 16 acoustic Low-Level Descriptors (LLDs) and their first delta regression. The LLDs are Mel-frequency cepstral coefficients 1-12, pitch frequency, harmonics-to-noise ratio by autocorrelation function, zero-crossing-rate, and root mean square of frame energy. The 12 functionals are minimum, maximum, mean, standard deviation, kurtosis, skewness, ranges, relative position, and two linear regression coefficients with their mean square error. Finally, the features are standardized for each corpus.

3.2. Distribution matching and data split

To match the feature distribution of the splitter, D_S, and emotion corpora, D_E, we train a Shared-Hidden-Layer Auto-encoder (SHLA). SHLA extracts the shared representation between the two corpora as well as reduces the dimensionality of the features. SHLA has been used to transfer knowledge between two corpora to achieve higher classification performance on unlabeled datasets [15], [18]. Having an emotion corpus $D_E^{m\times Q}$ and a splitter corpus $D_S^{n\times Q}$ with m and n samples and Q equivalent features, an artificial neural network with Q neurons in the input layer, H ($H < Q$) neurons in the hidden layer, and $2Q$ neurons in the output layer is created (See Figure 1). A gradient descent approach is performed to tune the weights. Finally, the outputs of the hidden layer ($\tilde{D_S}$ and $\tilde{D_E}$) represent the shared view of the two corpora. Once the SHLA is trained, we train a classifier on $\tilde{D_S}$ and apply it on $\tilde{D_E}$. By sorting the posterior probabilities and choosing the median as the splitting point, we divide the emotion corpora into two

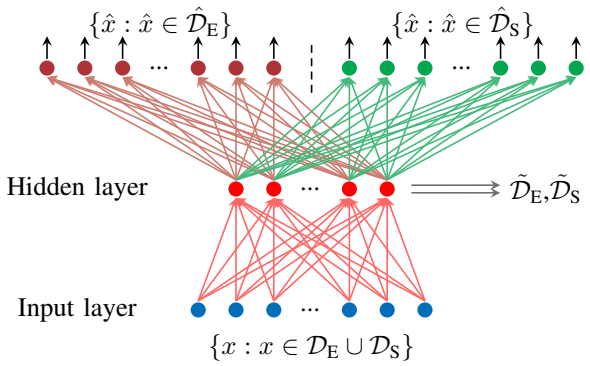

Figure 1. Structure of the shared-hidden-layer autoencoder (SHLA) on the Spliter set $\mathcal{D}^{\mathrm{S}}$ and Emotion set $\mathcal{D}^{\mathrm{E}}$. The SHLA shares same parameters for the mapping from the input layer to the hidden layer, but uses independent parameters for the corresponding reconstructions $\hat{\mathcal{D}}_{\mathrm{S}}$ and $\hat{\mathcal{D}}_{\mathrm{E}}$. After training, the shared representations $\tilde{\mathcal{D}}_{\mathrm{S}}$ and $\tilde{\mathcal{D}}_{\mathrm{E}}$ will be obtained from the hidden-layer.

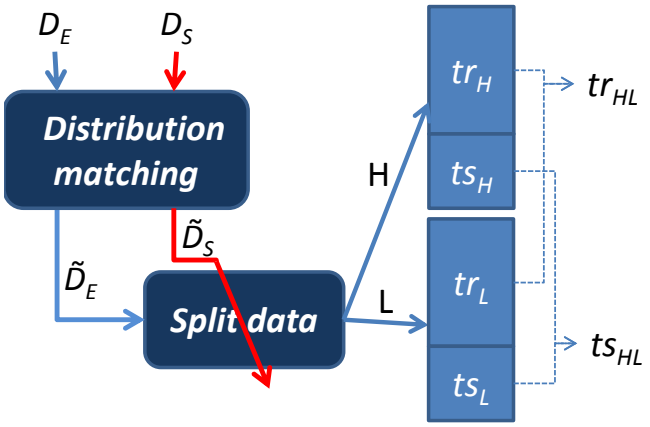

Figure 2. Data preparation schema. D_S and D_E are splitter and emotion speech corpora, respectively. $\tilde{D_S}$ and $\tilde{D_E}$ are the matched and reduced features of the splitter and emotion corpora. L and H correspond to Low and High scored personality traits, Male and Female, or Young and Adult.

equal-sized categories (representing low-high scored personality traits, young-old, or female-male). The schematic of this process is shown in Figure 2. Once the emotion corpus is split, we again split each category randomly into training (tr) and test (ts) sets for cross validation.

3.3. Analyses

We follow two distinct analyses. In the first analysis, we train a classifier on the training sets of the both categories (tr_{HL}) and test it (i) on the test sets of both categories (ts_{HL}) and (ii) on the test set of each category (ts_H and ts_L), separately.

In the second analysis, we train the classifier only on one category (e. g., tr_H) and test it on the same category (e. g., ts_H) or the contrary category (e. g., ts_L). This analysis helps to understand if training on a specific factor could benefit the overall recognition performance and if it affects the emotion recognition on the contrary category.

Finally, based on the results we can define sets of rules for the selection of the classifiers (out of the classifiers which are trained on tr_L, tr_H, or tr_{HL}) to achieve higher recognition performance.

All the analyses are performed with 30 iterations with repeated random sub-sampling, where 70% of the data is used for the training and the rest for the test. Support Vector Machines with linear kernel have been used as the classifier and the C parameter is optimized by cross-validation on the training set. Pair-wised two-sided t-test with $\alpha = 0.05$ has been applied to compare the mean accuracies.

3.4. Corpora

Two sets of speech corpora are used. The first set (splitters) is to train the models on age, gender, or personality traits to split the emotion corpora into two categories. The second set contains emotion corpora which are used for training and classification of emotional speech.

3.4.1. Splitter corpora. To split the emotion corpora according to the personality trait, we used the annotated dataset of Personality Sub-Challenge in Interspeech 2012 [19]. This corpus contains 640 clips (from 322 speakers) from French news bulletins of Radio Suisse Romande. The total length of the clips is about one hour and 40 minutes. Eleven judges performed the personality assessment based on Big-Five personality subscales. A binarized label is obtained if the majority of judges assign scores higher than their average for the same trait.

For splitting the emotion corpora based on the gender or age of the speakers, we used aGender database [20]. It contains 13985 male utterances, 14135 female utterances, and 4406 child utterances in the training set. The age of the speakers is between 7 and 80 years old. Note that, we excluded the child utterances (age $<$ 15 years old) from our analyses.

3.4.2. Emotion corpora. Four speech emotion corpora are experimented. Fau Aibo Emotion Corpus (Fau AEC) contains recordings of 51 children at the age between 10 to 13 years old interacting with Aibo robot in German [21]. The Audiovisual Interest Corpus (TUM-AVIC) consists of spontaneous speech and natural emotion and provides continuous labels for the level of interest [22]. The Speech Under Simulated and Actual Stress (SUSAS) corpus consists of 35 English air-commands in the speaker states high stress, medium stress, neutral, fear, and scream [23]. Finally, the eNTERFACE corpus consists of recordings of naïve subjects from 14 nations speaking predefined spoken content in English [24]. Particular emotion is elicited after listening to short stories.

Some information of these corpora are summarized in Table 1. Furthermore, we unified the labels by mapping them onto Positive and Negative valence as provided in Table 2. To design classifier selection rules, we used AVIC,

TABLE 1. SPECIFICATIONS OF THE CHOSEN EMOTIONAL SPEECH CORPORA.

Corpus	Age	Language	Speech	Emotion	# Valence -	# Valence +	# All	h:mm	#m	#f	Recording condition
TUM AVIC	adults	English	variable	natural	553	2 449	3 002	1:47	11	10	studio
eNTERFACE	adults	English	fixed	induced	855	422	1 277	1:00	34	8	normal
SUSAS	adults	English	fixed	natural	1 616	1 977	3 593	1:01	4	3	noisy
FAU AEC (Tr)	children	German	variable	natural	3 358	6 601	9 959	5:15	13	13	normal
FAU AEC (Ts)	children	German	variable	natural	2 465	5 792	8 257	4:05	17	8	normal

TABLE 2. EMOTION CATEGORIES MAPPING ONTO NEGATIVE AND POSITIVE VALENCE FOR SIX DATABASES.

Corpus	Negative	Positive
FAU AEC	angry, touchy, emphatic, reprimanding	motherese, joyful, neutral, rest
TUM AVIC	boredom	neutral, joyful
eNTERFACE	anger, disgust, fear, sadness	joy, surprise
SUSAS	high stress, screaming, fear	medium stress, neutral

eNTERFACE, SUSAS, and the training set of FAU AEC. To examine the efficiency of the designed rules, we train emotion classifiers on the training set of the FAU AEC and test them on its test set.

4. Results

The aggregated results over all four emotional speech corpora are provided in Table 3 (a). The cross validation is done for each factor separately, and therefore, analyses of the factors (columns) are independent from each other. The highest accuracy improvement (+2.66%) is achieved when a model is trained on Young speakers' utterances and tested on the same category. After that, the largest improvement is achieved when a model is trained and tested on Female speakers' utterances, or is trained on all ages and tested on Young speakers' utterances (+2.47%). It follows with the model which is trained on both Genders and tested on Female speakers (+1.63%). Then, there is a slight improvement when models are trained on C_{HL} and tested on C_L (+0.79%), trained on N_{HL} and tested on N_H (+0.75%), trained on C_L and tested on C_L (+0.68%), where C and N stand for Conscientiousness and Neuroticism.

Furthermore, there are performance degradations when we cross between contrary factor categories for the training and test data (e. g. N_L and N_H). The degradation is between -1.21% down to -10.28% (trained on E_H and tested on E_L). Similarly, there is about 6% degradation from Female to Male, Male to Female, and Young to Adult, as well as 8% degradation from Adult to Young. The results confirm that the age of the speaker has the highest impact on speech emotion recognition followed by the gender of the speakers and their personality trait.

4.1. Model Selection

The results help to select appropriate models while classifying emotions. For example, when a speaker is detected as Young, instead of a model which is trained on data from both Young and Adult utterances, we should choose a model which is trained only on Young speakers' utterances. Similarly, if the speaker is detected as Adult, a model which is trained on both Young and Adult speakers is more appropriate. Therefore, considering only the significant improvements, we set four simple rules based on the factors as follows:

1) If $Age == Young$, then
   use M_{Young}, #you may gain 2.66 UAR
   otherwise use M_{all}. #at most you may lose -4.63 UAR
2) If $Gender == Female$, then
   use M_{Female}, #you may gain 2.47 UAR
   otherwise use M_{all}. #at most you may lose -2.78 UAR
3) If $Conscientiousness == Low$, then
   use M_{all}, #you may gain 0.79 UAR
   otherwise use M_{High}.#at most you may lose -2.85 UAR
4) If $Neuroticism == High$, then
   use M_{All}, #you may gain 0.75 UAR
   otherwise use M_{Low}.#at most you may lose -3.34 UAR

where M_{all} is the model which is trained on all data, M_{Young} is the model which is trained on Young speaker utterances and so on. Note that, in this study, we consider these rules independent of each other and we investigate the results for each rule separately. Designing more complicated rules which uses all the criteria (such as designing a decision tree) is also a possible approach and beyond the scope of this paper. Table 3 (b) shows the results on the test part of the FAU AEC dataset by applying the above-mentioned rules. There is 1.04%, 0.40%, and 0.39% significant improvement by applying the 3rd, the 4th, and the 1st rules, respectively. Applying the 2nd rule does not yield significant improvement. However, all the rules yield positive improvement. This results confirm the benefit of data or model selection for a specific speaker to achieve higher emotion recognition accuracy.

5. Discussion

Although there has been a wide range of research on automatic emotion recognition systems, the effects of human factors on these systems are not studied yet. In this paper, we have evaluated how Big-Five personality traits, age, and gender can affect an emotion recognition system. We found that,

TABLE 3. (A) RECOGNITION PERFORMANCE (UAR) OVER ALL THE EMOTIONAL SPEECH CORPORA. (B) RECOGNITION PERFORMANCE AFTER RULE-BASED RECOGNIZER SELECTION ON FAU AEC (TS). THE HIGHEST ACCURACY ON EACH COLUMN IS BOLD-FACED. NOT SIGNIFICANT VALUES ARE SUPERSCRIPTED BY ^{ns}.

(a)

	Train	Test	O	C	E	A	N	Gender	Age
UAR	tr_{HL}	ts_{HL}	**73.66**	73.78	**73.33**	**73.66**	73.66	$(tr_{MF} \rightarrow ts_{MF})$ 73.38	$(tr_{AY} \rightarrow ts_{AY})$ 74.11
ΔUAR	tr_{HL}	ts_H	0.07^{ns}	-3.03	0.11^{ns}	-0.77	**0.75**	$(tr_{MF} \rightarrow ts_F)$ 1.63	$(tr_{AY} \rightarrow ts_A)$ -4.63
	tr_{HL}	ts_L	-1.83	**0.79**	-3.47	-1.21	-3.66	$(tr_{MF} \rightarrow ts_M)$ -2.78	$(tr_{AY} \rightarrow ts_Y)$ 2.47
ΔUAR	tr_H	ts_H	0.00^{ns}	-2.85	-0.02^{ns}	-0.24^{ns}	0.51^{ns}	$(tr_F \rightarrow ts_F)$ **2.47**	$(tr_A \rightarrow ts_A)$ -5.07
	tr_L	ts_L	-2.23	0.68	-2.61	-1.33	-3.34	$(tr_M \rightarrow ts_M)$ -3.15	$(tr_Y \rightarrow ts_Y)$ **2.66**
	tr_H	ts_L	-5.39	-4.61	-10.28	-8.84	-9.88	$(tr_F \rightarrow ts_M)$ -6.59	$(tr_A \rightarrow ts_Y)$ -8.10
	tr_L	ts_H	-7.19	-6.74	-6.78	-2.88	-4.21	$(tr_M \rightarrow ts_M)$ -5.65	$(tr_Y \rightarrow ts_A)$ -6.39

(b)

UAR	tr_{HL}	ts_{HL}	60.94	61.15	60.99	60.86	60.93	$(tr_{MF} \rightarrow ts_{MF})$ 63.16	$(tr_{AY} \rightarrow ts_{AY})$ 62.41
ΔUAR	Rule		-	1.04	-	-	0.40	0.18^{ns}	0.39

the age of the speaker is the most important factor: Young speakers' emotions are highly discriminant. Then, females' emotions are the second most important factor. These results are congruent with the findings of [13] and [8], where they showed that females and young people are more emotionally expressive. Moreover, Conscientiousness and Neuroticism are the other involved factors. High Neuroticism and low Conscientiousness improve the recognition performance and these are also congruent with [12] and [7] where they found Neuroticism is directly and Conscientiousness is inversely related to negative emotion expressivity.

Moreover, comparing the models which are trained on tr_{HL} and tested on high scored personality traits (ts_H), we see that high Neuroticism, high Agreeableness, high Extroversion, and high Openness convey higher performance than their corresponding low scored traits (ts_L). This is also inline with [7].

Additionally, we found that cross-factor-category classification always decreases the performance. This finding is closely related to [9] where they found that in-group emotion expression and recognition has higher accuracy.

Nevertheless, we should take the results with a grain of salt. In this study different databases are used and a distribution matching is applied to reduce data dissimilarity. Both add noises to our data-driven analyses. An ideal case would be to analyze a dataset, in which, personality, age, gender, and emotions are all labeled.

6. Conclusion

In this paper, we investigated the effect of different human factors (personality trait, age, and gender) on the performance of speech emotion recognition system. High improvement has been achieved when the system is trained and tested on Young or Female speakers. Additionally, high Neuroticism and low Conscientiousness improve the accuracy. The results are in line with the psychological studies that target the effect of individual factors on emotion expressivity. Therefore, the performance of a speech emotion recognizer is closely correlated with the emotion expressivity of the individuals whose speech are used for training and applying the recognition models. We also found that cross-factor classification will significantly decrease the recognition performance. Based on these findings, we set some rules for data and model selection and we could increase the recognition performance on another emotional speech database.

Acknowledgement

The research leading to these result has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation programme and EFPIA.

References

[1] S. Berkovsky, "Emotion-based movie recommendations: How far can we take this?" in *Proc. 3rd Workshop on Emotions and Personality in Personalized Systems*. Vienna, Austria: ACM, 2015, pp. 1–1.

[2] I. Jraidi, "Modélisation des émotions de lapprenant et interventions implicites pour les Systèmes Tutoriels Intelligents," Ph.D. dissertation, Université de Montréal, 2008.

[3] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. De Sevin, M. Valstar, and M. Wöllmer, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.

[4] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[5] H. Sagha, E. Coutinho, and B. Schuller, "Exploring the importance of individual differences to the automatic estimation of emotions induced by music," in *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*. Brisbane, Australia: ACM, 2015, pp. 57–63.

[6] J. J. Gross, O. P. John, and J. M. Richards, "The dissociation of emotion expression from emotion experience: A personality perspective," *Personality and Social Psychology Bulletin*, vol. 26, no. 6, pp. 712–726, 2000.

[7] J. J. Gross and O. P. John, "Facets of emotional expressivity: Three self-report factors and their correlates," *Personality and Individual Differences*, vol. 19, no. 4, pp. 555–568, 1995.

[8] A. M. Kring and A. H. Gordon, "Sex differences in emotion: expression, experience, and physiology," *Journal of personality and social psychology*, vol. 74, no. 3, pp. 686–703, 1998.

[9] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, pp. 203–235, 2002.

[10] J. J. Gross and O. P. John, "Revealing feelings: facets of emotional expressivity in self-reports, peer ratings, and behavior," *Journal of personality and social psychology*, vol. 72, no. 2, pp. 435–448, 1997.

[11] H. R. Riggio and R. E. Riggio, "Emotional expressiveness, extraversion, and neuroticism: A meta-analysis," *Journal of Nonverbal Behavior*, vol. 26, no. 4, pp. 195–218, 2002.

[12] J. A. A. Abe and C. E. Izard, "A longitudinal study of emotion expression and personality relations in early development," *Journal of personality and social psychology*, vol. 77, no. 3, pp. 566–577, 1999.

[13] J. J. Gross, L. L. Carstensen, M. Pasupathi, J. Tsai, C. Götestam Skorpen, and A. Y. Hsu, "Emotion and aging: experience, expression, and control," *Psychology and aging*, vol. 12, no. 4, pp. 590–599, 1997.

[14] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, Xian, China, Sept 2015, pp. 125–131.

[15] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Proc. 41st International Conference on Acoustics, Speech, and Signal Processing*. Shanghai, P. R. China: IEEE, March 2016, pp. 5800–5804.

[16] H. Sagha, P. Matejka, M. Gavryukova, F. Povolny, E. Marchi, and B. Schuller, "Enhancing multilingual recognition of emotion in speech by language identification," in *Proc. 17th Annual Conference of the International Speech Communication Association*. San Francisco, CA: ISCA, September 2016, pp. 2949–2953.

[17] F. Eyben and B. Schuller, "openSMILE:): The Munich open-source large-scale multimedia feature extractor," *SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

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

[19] B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. 13th Annual Conference of the International Speech Communication Association*, Portland, OR, USA, Sep 2012, pp. 254–257.

[20] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech." in *Proc. 7th International Conference of Language Resources and Evaluation*, Malta, 2010.

[21] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. Logos Verlag, Berlin, 2009.

[22] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[23] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database," in *Proc. 5th European Conference on Speech Communication and Technology*, vol. 97, no. 4, 1997, pp. 1743–46.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd International Conference on Data Engineering Workshops*. IEEE, 2006, pp. 8–8.