Enhancing multilingual recognition of emotion in speech by language identification

Hesam Sagha[1], *Pavel Matejka*[2,3], *Maryna Gavryukova*[1], *Filip Povolny*[2], *Erik Marchi*[1], *Björn Schuller* [1,4]

[1]Chair of Complex & Intelligent Systems, University of Passau. Germany
[2]Phonexia Brno, Czech Republic
[3]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic
[4]Department of Computing, Imperial College London, UK

hesam.sagha@uni-passau.de, matejka@phonexia.com

Abstract

We investigate, for the first time, if applying model selection based on automatic language identification (LID) can improve multilingual recognition of emotion in speech. Six emotional speech corpora from three language families (Germanic, Romance, Sino-Tibetan) are evaluated. The emotions are represented by the quadrants in the arousal/valence plane, i. e., positive/negative arousal/valence. Four selection approaches for choosing an optimal training set depending on the current language are compared: within the same language family, across language family, use of all available corpora, and selection based on the automatic LID. We found that, on average, the proposed LID approach for selecting training corpora is superior to using all the available corpora when the spoken language is not known.

Index Terms: multilingual emotion recognition, language identification, language families

1. Introduction

Each culture imposes certain attitudes, behaviors, verbal and non-verbal reactions different from other cultures among its individuals. Similarly, it influences emotion expression and perception. These variations affect cross cultural emotion comprehension by humans; Elfenbein and Ambady found that when emotions are expressed and recognized by the people of the same ethnic or regional group the emotion recognition accuracy is higher [1]. Equivalently, we expect that due to these variations, without knowing the cultural, ethnic or language background of a person, *automatic* emotion recognition (AER) is a difficult task and can be error-prone.

In the last decade, AER (especially from speech) has gained increasing attention in various domains, such as, health care [2], education [3], serious games [4], and robotics [5]. Despite decent performance being reported in research papers under laboratory conditions [6], emotion recognition from speech under real-life conditions still remains challenging; in particular, when considering the contextual dependencies of affective expressions across different speakers, languages and cultures.

To understand the features leading to these variations, a number of interdisciplinary studies paid attention to unique speech attributes within and across cultures, reporting strong universal similarities as well as cultural diversities (e. g., [7, 8, 9]). Likewise, Scherer et al. concluded that culture- and language-specific paralinguistic patterns may influence the emotion perception [10]. Furthermore, Feraru et al. investigated emotion recognition from speech on cross-language families by including less researched languages from completly different language families such as Burmese, Romanian or Turkish [11]. They found that, AER for corpora of the same language has the highest accuracy while emotion recognition across language families has the lowest. In the middle, in terms of accuracy, came emotion recognition within the same language family. Therefore, we can infer that speech linguistic features carry information about the culture and the way emotions are expressed or perceived. Consequently, these inherent linguistic features in speech could be used to enhance cross-lingual (-cultural) emotion recognition.

In this paper, we investigate if applying a model selection technique based on language identification (LID) for multilingual emotion recognition could improve speech emotion recognition (SER) accuracy. The databases given in this study represent three different language families (Germanic, Romance, Sino-Tibetan). We compare four selected approaches for choosing an optimal training set depending on the current language: i) a supervised model selection (where the language of the utterance is known), ii) cross-family model selection, iii) using a model which is trained on all available corpora, and iv) selection based on automatic LID.

The paper is organized as follows: the next section briefly reviews the literature on multilingual emotion recognition. In Sections 3 and 4, we describe our approach and the designed experiments, respectively. In Section 5, the results will be provided and finally, in Section 6, we draw conclusions and suggest future work.

2. Literature review

Although there is extensive research on enhancing emotion and sentiment recognition from multilingual text [12, 13], there is cosiderably less effort on the emotion analysis from multilingual speech. This is mostly due to the lack of multilingual databases with equal conditions as well as the assumption that paralinguistic features (i. e., how something is said) represent emotions in speech more than linguistic features (i. e., what is said). There exists a plethora of databases considering emotional state [14]. However, these resources are existing mostly for all Indo-European languages, e. g. English, French or German [15]. Nevertheless, one of the only multilingual emotional speech database is INTERFACE [16] which is not freely available. Despite this lack of resources, a preliminary study shows that within English, Hebrew, and Swedish speech samples, 'Afraid' is harder to be recognize in Hebrew language,

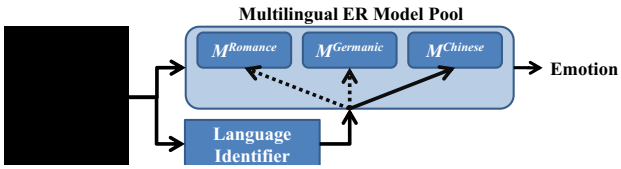

Figure 1: Schematic of the proposed multilingual AER.

while 'Anger' is harder in English and Swedish [4]. Furthermore, transfer learning has been applied to recognize emotions from multilingual speech. This approach allows the use of knowledge obtained from other databases and to be transfered for use as knowledge for a new database. Shared-Hidden-Layer Auto-Encoders [17] and Canonical Correlation Analysis [18] are proposed to match the feature distribution between multilingual emotional speech corpora. To the best of the authors knowledge, this is the first time that one uses language identification to select an appropriate model for emotion recognition.

3. Method

Our approach to the multilingual emotion analysis is to first train a model for each target language. Then, for a new utterance, we detect the language using a language identifier and select the corresponding model for emotion recognition. The schema of this process is shown in Fig. 1. For the remainder of this section, we describe the language identifier and the speech emotion recognizer.

3.1. LID

Our language identification is based on the i-vector approach with bottleneck features and a Gaussian linear backend as classifier [19]. The i-vector approach was introduced in speaker recognition [20], but has been widely used in multiple fields of speech processing, such as age estimation [21], emotion detection [22], depression analysis [23], speech recognition [24], and also language recognition [25]. Since then, the major improvement comes from the use of bottleneck features to reduce the error rates of the language recognition system by 50 % relative with respect to the conventionally used shifted delta cepstra [19]. Our system consists of 5 blocks:

Voice activity detection Small neural network trained with 2 outputs (one for speech and second for non-speech) is used for selecting only the speech parts of the recording.

Bottleneck Feature Extraction Bottleneck Neural-Network (BN-NN) refers to a topology of a NN, whose hidden layers has lower dimensionality than the surrounding layers. A bottleneck feature vector is generally understood as a by-product of forwarding a primary input feature vector through the BN-NN and reading off the vector of values at the bottleneck layer. We have used a cascade of two such NNs for our experiments. The output of the first network is *stacked* in time, defining context-dependent input features for the second NN, hence the term Stacked Bottleneck Features.

The NN input features are 24 log Mel-scale filter bank outputs augmented with fundamental frequency and probability of voicing features based on [26]. The conversation-side based mean subtraction is applied on the whole feature vector. 11 frames of log filter bank outputs and fundamental frequency features are stacked together. Hamming window followed by discrete cosine transform consisting of the 0^{th} to 5^{th} base are applied on the time trajectory of each parameter resulting in $(24+2)\times 6 = 156$ coefficients on the first stage NN input.

The configuration for the first NN is $156 \times D_H \times D_H \times D_{BN} \times D_H \times K$, where K is the number of targets (phoneme states from 5 languages in our case). The dimensionality of the bottleneck layer, D_{BN} was fixed to 80. This was shown as optimal in [27]. The dimensionality of other hidden layers was set to 1500. The bottleneck outputs from the first NN are sampled at times $t-10$, $t-5$, t, $t+5$, and $t+10$, where t is the index of the current frame. The resulting 400-dimensional features are input to the second stage NN with the same topology as first stage. The 80 bottleneck outputs from the second NN (referred to SBN) are taken as features for the conventional GMM/UBM i-vector based SID system.

For training the neural networks, the IARPA Babel Program data[1] were used. This data simulates a case of what one could collect in limited time from a completely new language. It consists mainly of telephone conversational speech, but scripted recordings as well as far field recordings are also present. We have used first five languages from the collection (Cantonese, Pashto, Turkish, Tagalog, Vietnamese). For more analysis about multilingual SBN see [19].

i-vector provides an elegant way of reducing large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information. The technique was originally inspired by the Joint Factor Analysis (JFA) framework introduced in [28].

The main idea is that the utterance-dependent Gaussian Mixture Model (GMM) supervector of concatenated GMM mean vectors $\mathbf{s}$ can be modeled as:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}$ is the Universal Background Model (UBM) GMM mean supervector, $\mathbf{T}$ is a low-rank matrix representing M bases spanning subspace with important variability in the mean supervector space, and $\mathbf{w}$ is a latent variable of size M with standard normal distribution.

For each observation $\mathcal{X}$, the aim is to compute the parameters of the posterior probability of $\mathbf{w}$:

$$\mathrm{p}(\mathbf{w}|\mathcal{X}) = \mathcal{N}(\mathbf{w}; \mathbf{w}_{\mathcal{X}}, \mathbf{L}_{\mathcal{X}}^{-1}). \tag{2}$$

The i-vector $\boldsymbol{\phi}$ is the Maximum a Posteriori (MAP) point estimate of the variable $\mathbf{w}$, i. e., the mean $\mathbf{w}_{\mathcal{X}}$ of the posterior distribution $\mathrm{p}(\mathbf{w}|\mathcal{X})$. It maps most of the relevant information from a variable-length observation $\mathcal{X}$ to a fixed- (small-) dimensional vector. $\mathbf{L}_{\mathcal{X}}$ is the precision of the posterior distribution.

We used 2 048 Gaussian Mixture Components with diagonal covariances and an i-vector with 600 dimensions.

Gaussian Linear Classifier Utterance level score vectors (e. g., i-vectors) are modeled as having a shared within-class covariance and language dependent means. These parameters are estimated with maximum-likelihood. The scores are computed from the constant and linear terms of the Gaussian log-likelihoods for each language. The data for training the language models come from the past NIST Language recognition evaluations[2]. We used 8 models: Chinese Dialects, Mandarin Chinese, American English, British English, Indian English, German, French, and Spanish.

[1] Collected by Appen, http://www.appenbutlerhill.com
[2] NIST LRE: lre.nist.gov

Table 1: Corpora information. (#m): number of male speaker, (#f): number of female speakers.

Corpus	Language	language family	#m	#f	#instances	avg utterance length (s)
EU-EmoSS	English	Germanic	19	17	1534	2 .08
EU-EmoSS	German	Germanic	2	4	258	1.90
EU-EmoSS	French	Romance	3	5	375	1.86
EU-EmoSS	Spanish	Romance	3	3	252	1.71
VESD	Mandarin Chinese	Sino-Tibetan	2	2	874	1.66
CASIA	Mandarin Chinese	Sino-Tibetan	2	2	1200	1.91

Table 2: Mapping of class labels onto Negative/Positive Arousal/valence. Acronyms: Neu(tral), Sur(prised), Ang(er), Hap(py), Fru(strated), Int(erested), Dis(appointed), Unf(riendly), Disg(ust), Unfr(iendly)

Corpus	Negative Arousal	(#)	Positive Arousal	(#)	Negative Valence	(#)	Positive Valence	(#)
EmoSS (EN)	Sad, Afraid, Disg.,	(768)	Hap., Ang., Sur.,	(748)	Sad, Ang., Afraid, Disg.,	(892)	Happy, Sur.,	(642)
EmoSS (DE)	Worried, Bored, Unfr,	(147)	Joking, Int.,	(111)	Worried, Bored, Hurt,	(160)	Joking, Int.,	(98)
EmoSS (FR)	Neu., Sneaky, Jealous,	(215)	Proud, Hurt,	(160)	Sneaky, Jealous, Unfr.,	(234)	Proud, Kind,	(141)
EmoSS (ES)	Ashamed, Dis., Fru.	(144)	Kind, Excited	(108)	Ashamed, Dis., Fru.	(156)	Neu., Excited	(96)
VESD	Neu., Sadness, Disg.	(387)	Hap., Sur., Fear, Ang.	(445)	Ang., Sad, Fear, Disg.	(551)	Neu., Hap., Sur.	(321)
CASIA	Neu., Sadness	(800)	Hap., Sur., Fear, Ang.	(400)	Ang., Sad, Fear	(600)	Neu., Hap., Sur.	(600)

Table 3: Four approaches for selecting training corpora (An example for an utterance in the test Corpus 1). X: chosen, ?: selection based.

	Family 1		Family 2	
	Corp. 1	Corp. 2	Corp. 3	Corp. 4
Same family		X		
Cross family			X	X
All corpora		X	X	X
Selective (LID)		?	?	?

3.2. Speech Emotion Recognition

From each utterance, 384 features are extracted using openSMILE [29]. The feature set was introduced in the Interspeech 2009 Emotion Challenge; it contains 12 functionals of 2 $\times$ 16 acoustic Low-Level Descriptors (LLDs) including their first delta regression. The LLDs are zero-crossing rate, root mean square of frame energy, pitch frequency, harmonics-to-noise ratio by autocorrelation function and Mel-frequency cepstral coefficients 1–12. The 12 functionals are minimum, maximum, mean, standard deviation, kurtosis, skewness, relative position, ranges, and two linear regression coefficients with their mean square error. Additionally, we normalized each corpus to have zero mean and standard deviation of one. It has been shown that, the corpus-normalization has positive effects on cross-corpus emotion recognition [30]. Finally, as the classifier, we used Support Vector Machines (SVM) with linear kernel.

4. Experiment

The experimental setup is as follows. We select two separate corpora for each language: One for training a speech emotion recognizer, and one for the test. Having these two corpora with the same recording conditions is ideal. However, due to the lack of multilingual speech emotion databases, we opt for having two corpora for each language family. We compare four emotion model selection approaches for each utterance: the same language family, across language family, all languages, and selective language family (see Table 3). For the *same language family* we select a model which is trained on the other corpus within the same language family while for the *across language family* a model which is trained on both corpora in the other language family is selected. In the case of *all languages*, we select a model which is trained on all other available corpora. Finally, for the *selective language family* we select the model based on the identified language family: trained on the other corpus in the same family or on the two corpora of the other language family. To have a fair comparison, for each case, we upsample the training set in such a way to have the same number of instances as the whole datapoints in all corpora as well as to have the same number of instances for each class. In the next section, we describe the corpora for our analyses.

Databases

Six databases with languages from different language families (Romance, Germanic, Sino-Tibetan) have been evaluated in this paper. Table 1 gives an overview of the selected databases. The Chinese Vocal emotional stimuli Database (VESD) is recorded in Mandarin Chinese. Thirty five pseudo sentences were selected and read by 4 subjects (2 males and 2 females, with a mean age of 24.3 years) to express 7 emotion states, namely anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality [31]. The Chinese emotional speech corpus CASIA consists of 5 emotions (angry, fear, happy, sad, and neutral). For each emotion, 500 sentences were read by 4 professional speakers (2 males and 2 females) [32]. The EU-Emotion Stimulus Set (EU-EMOSS) [33] is a newly developed collection of dynamic multi-modal emotion (facial expressions, voice and body gesture) and mental state representations. A total of 20 emotions and mental states (afraid, angry, ashamed, bored, disappointed, disgusted, excited, frustrated, happy, hurt, interested, jealous, joking, kind, proud, sad, sneaky, surprised, unfriendly plus neutral) are represented there. This emotion set is portrayed by a multi-ethnic (German, French, Spanish, and English) group of child and adult actors aged 10–70 years old (ten female and nine male). The database was recorded in the context of the ASC-Inclusion project [2]. To have unified emotion labels for each databases, we mapped the labels onto 4 classes positive/negative Arousal/Valence. These mappings are provided in the Table 2.

Table 5: Language identification and emotion recognition accuracy (Unweighted Average Recall). On diagonals, each language is tested on a model which is trained on its pair within the same family. On off diagonals, that language is tested on a model which is trained on other language families (stacking datasets). The LID column is the accuracy after language identification and selecting the corresponding model. The highest accuracies for each test corpus and emotion dimension are bold-faced.

Family	Database	LID	Train set									
			Arousal					Valence				
		ACC	Germ.	Sin.-T.	Rom.	All	LID	Germ.	Sin.-T.	Rom.	All	LID
Germanic	EMOSS English	58.9	63.6	64.6	**65.6**	65.2	65.2	60.0	57.2	**60.2**	59.7	59.5
	EMOSS German	59.7	66.0	65.7	66.1	66.7	**66.8**	**64.3**	55.0	61.3	64.1	62.1
Sino-Tibetan	Chinese VESD	84.2	69.8	**80.0**	67.7	74.0	78.1	53.6	**63.4**	53.4	55.8	62.3
	Chinese CASIA	96.6	67.6	73.3	65.7	69.3	**73.6**	61.0	65.7	54.8	62.6	**65.8**
Romance	EMOSS French	69.3	66.2	65.6	**68.3**	66.0	67.7	63.1	52.1	**63.4**	58.8	60.8
	EMOSS Spanish	68.7	69.0	68.1	71.9	72.2	**73.0**	64.1	51.8	**66.8**	62.8	64.1

Table 4: LID family-wise confusion matrix (%)

Family	Germanic	Sino-Tibetan	Romance
Germanic	**59.0**	27.0	14.0
Sino-Tibetan	5.6	**91.4**	3.0
Romance	15.6	15.3	**69.1**

5. Result

The confution matrix of the language family identification is provided in Table 4. The accuracy on the Germanic languages is quite low; we speculate this is due to the large amount of short utterances (95% of the utterances have less than three seconds). Moreover, the accuracy is the highest for the Sino-Tibetan language family. This could be due to the high differentiation between Indo-European and Sino-Tibetan language families.

The emotion recognition results are provided in Table 5. Each row of the table indicates the test corpora. The 'LID ACC' column represents how accurate the LID could classify each corpus into the correct language family (3-classes). The training emotion corpora are either (i) the other language within the same language family (diagonals), (ii) the aggregation of the corpora in the other language family (off-diagonals), or (iii) both within and across language families ('All' column). As indicated in [11], we expect to have high accuracies on diagonals (i. e., same language family) with respect to the off-diagonals (i. e., cross language family). This is valid for the Sino-Tibetan and Romance families. However, with these corpora, the arousal dimension of Germanic languages, and valence dimension of the English language are slightly better recognized with their other Indo-European family corpora. Moreover, selecting all the available datasets does not improve the accuracy with respect to cross corpus analysis, except for the arousal dimension of the Spanish and German languages. This could be because of the noise added to the model through differences of emotion expressions between languages (cultures).

The 'LID' columns hold the accuracies using the proposed model selection technique based on the identified language family. If LID accuracy is 100%, we expect to have the same values for the LID column as in the diagonals. However, there could be a benefit in having imperfect LID; if LID incorrectly identifies the language of an utterance, probably that utterance is more similar to the detected language, rather than the actual language, and the model of the identified language could classify better that utterance. This can be perceived better as a clustering of the feature space using meta-data. Because of this imperfection, we gain slightly higher accuracy on the arousal dimension for German, Chinese (CASIA), and Spanish and on the valence dimension for Chinese (CASIA) databases.

Furthermore, in nine cases out of twelve, using LID outperforms using all the available datasets. This implies that targeting to a specific model is superior to a general model and adding extra data does not necessarily bring useful information.

6. Discussion and conclusion

Although emotions in speech are more perceivable through the paralinguistic features rather than the linguistics, the latter could bring some useful information on the cultural background of the speaker. This background defines certain productions and perceptions of the emotion which could be different from other cultures. Therefore, applying this knowledge to an emotion recognition system could be beneficial. As our results indicate, identifying the language of a speaker to some extent brings such knowledge, and by selecting an appropriate model based on that knowledge we could enhance the performance of the emotion recognition system.

Further, we found that, to recognize the emotions of a speaker whose language is unknown, it would be beneficial to use a language identifier and model selection instead of using a model which is trained based on all available languages.

Clearly, the results should be interpreted with utmost care and shall serve as tendencies due to the unavailability of multilingual speech emotion corpora with equal conditions. Additionally, most of the available emotional speech datasets contain very short utterances (less than 3 seconds) and cause high misclassification rates for the language identifier.

Obviously, these experiments should be redone with more languages or language families. Moreover, we would like to compare the proposed approach with adaptation techniques without knowing the spoken language.

7. Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 2020 Programme research and innovation programme under grant agreements No. 644632 (MixedEmotions) and IMI RADAR-CNS under grant agreement No. 115902. The work was further by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", and the Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

8. References

[1] 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.

[2] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle *et al.*, "Recent developments and results of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions," in *Proc. 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI)*. Atlanta, GA: ACM, March 2015, 9 pages.

[3] R. A. Calvo and S. D'Mello, "Frontiers of affect-aware learning technologies," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 86–89, Nov. 2012.

[4] E. Marchi, B. Schuller, S. Baron-Cohen, A. Lassalle *et al.*, "Voice Emotion Games: Language and Emotion in the Voice of Children with Autism Spectrum Condition," in *Proc. 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI)*. Atlanta, GA: ACM, March 2015.

[5] E. Marchi, F. Ringeval, and B. Schuller, "Voice-enabled assistive robots for handling autism spectrum conditions: an examination of the role of prosody," in *Speech and Automata in Health Care (Speech Technology and Text Mining in Medicine and Healthcare)*, A. Neustein, Ed. Boston/Berlin/Munich: De Gruyter, 2014, pp. 207–236.

[6] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. 11th Biannual IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Merano, Italy: IEEE, Dec. 2009, pp. 552–557.

[7] T. Shochi, A. Rilliard, V. Aubergé, and D. Erickson, "Intercultural Perception of English, French and Japanese Social Affective Prosody," *The role of prosody in Affective Speech*, vol. 97, p. 31, 2009.

[8] G. A. Bryant and H. C. Barrett, "Vocal emotion recognition across disparate cultures," *Journal of Cognition and Culture*, vol. 8, no. 1, pp. 135–148, 2008.

[9] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Bölte, P. Arora, and R. Häb-Umbach, "Typicality and Emotion in the Voice of Children with Autism Spectrum Condition: Evidence Across Three Languages," in *Proc. Interspeech*. Dresden, Germany: ISCA, Sep. 2015, pp. 115–119.

[10] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology*, vol. 32, no. 1, pp. 76–92, 2001.

[11] S. Feraru, D. Schuller, and B. Schuller, "Cross-Language Acoustic Emotion Recognition: An Overview and Some Tendencies," in *Proc. 6th biannual Conference on Affective Computing and Intelligent Interaction*, AAAC. Xi'an, P. R. China: IEEE, Sep. 2015, pp. 125–131.

[12] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: are more languages better?" in *Proc. 23rd International Conference on Computational Linguistics*, Beijing, China, August 2010, pp. 28–36.

[13] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis," in *Proc. 13th International Conference on Multimodal Interfaces*. New York, USA: ACM, Nov. 2011, p. 169.

[14] D. Ververidis and C. Kotropoulos, "A review of emotional speech databases," in *Proc. Panhellenic Conference on Informatics*, Thessaloniki, Greece, 2003, pp. 560–574.

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

[16] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface Databases: Design and Collection of a Multilingual Emotional Speech Database," in *Proc. 3rd internationl conference on language resources and evaluation*, Las Palmas de Gran Canaria, 2002, pp. 2024–2028.

[17] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse Autoencoder-based Feature Transfer Learning for Speech Emotion Recognition," in *Proc. 5th biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, HUMAINE Association. Geneva, Switzerland: IEEE, Sep. 2013, pp. 511–516.

[18] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Shanghai, P. R. China: IEEE, March 2016, pp. 5800–5804.

[19] R. Fér, P. Matějka, F. Grézl, O. Plchot, and J. Černocký, "Multilingual bottleneck features for language recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 389–393.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[21] A. Fedorova, O. Glembek, P. Matejka, and T. Kinnunen, "Exploring ANN back-ends for i-vector based speaker age estimation," in *Proc. Interspeech*, Dresden, Germany, 2015.

[22] M. Kockmann, L. Burget, and J. Černocký, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition," *Speech Communication*, vol. 53, no. 9, pp. 1172–1185, 2011.

[23] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy: IEEE, 2014, pp. 970–974.

[24] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černocký, "iVector-based discriminative adaptation for automatic speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Hawaii, US: IEEE Signal Processing Society, 2011, pp. 152–157.

[25] D. G. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Proc. Interspeech*, no. 8. International Speech Communication Association, 2011, pp. 861–864.

[26] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. New York: Elseviever, 1995.

[27] P. Matějka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop Odyssey*, no. 6. International Speech Communication Association, 2014, pp. 299–304.

[28] P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[29] 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.

[30] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, 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.

[31] P. Liu and M. D. Pell, "Recognizing vocal emotions in mandarin chinese: A validated database of chinese vocal emotional stimuli," *Behavior research methods*, vol. 44, no. 4, pp. 1042–1051, 2012.

[32] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *Proc. International Conference on Information Engineering and Computer Science*. IEEE, 2009, pp. 1–4.

[33] H. OReilly, D. Pigat, S. Fridenson, S. Berggren, S. Tal, O. Golan, S. Bölte, S. Baron-Cohen, and D. Lundqvist, "The EU-Emotion Stimulus Set: A validation study," *Behavior research methods*, vol. 48, no. 2, pp. 1–10, 2015.