Cross-Language Acoustic Emotion Recognition: An Overview and Some Tendencies

Silvia Monica Feraru[1], Dagmar Schuller[2], and Björn Schuller[1,2,3,4]
[1]Machine Intelligence & Signal Processing group, MMK, Technische Universität München, 80333 Munich, Germany
[2]audEERING UG, Landsberger Strasse 46D, 82205 Gilching, Germany
[3]Chair of Complex & Intelligent Systems, University of Passau, Innstrasse 43, 94032 Passau, Germany
[4]Department of Computing, Imperial College London, 180 Queen's Gate, SW7 2AZ London, UK
Email: schuller@ieee.org

***Abstract*—Automatic emotion recognition from speech has matured close to the point where it reaches broader commercial interest. One of the last major limiting factors is the ability to deal with multilingual inputs as will be given in a real-life operating system in many if not most cases. As in real-life scenarios speech is often used mixed across languages more experience will be needed in performance effects of cross-language recognition. In this contribution we first provide an overview on languages covered in the research on emotion and speech finding that only roughly two thirds of native speakers' languages are so far touched upon. We thus next shed light on mis-matched vs matched condition emotion recognition across a variety of languages. By intention, we include less researched languages of more distant language families such as Burmese, Romanian or Turkish. Binary arousal and valence mapping is employed in order to be able to train and test across databases that have originally been labelled in diverse categories. In the result – as one may expect – arousal recognition works considerably better across languages than valence, and cross-language recognition falls considerably behind within-language recognition. However, within-language family recognition seems to provide an 'emergency-solution' in case of missing language resources, and the observed notable differences depending on the combination of languages show a number of interesting effects.**

Keywords: ***Speech Emotion Recognition***; ***Multilinguality***; ***Cross-Corpus***

I. Introduction

Automatic emotion recognition from speech has matured close to the point where it reaches broader commercial interest. This is shown, for example, by the recent emergence of start-ups focussing on the task such as audEERING, Beyond Verbal, emospeech, nexidia, and previously Nemesysco. One of the last major limiting factors at this point thus becomes the ability to ensure coping with different languages and to deal also with multilingual inputs as will be given in a real-life speech-based emotion recognition operating system in many if not most cases. As in real-world scenarios speech is often used mixed across languages such as when using, e. g., English expressions in one's own (non-English) language, more experience will be needed in performance effects of cross-language recognition.

A number of studies has investigated the effects of multi- and cross-language human emotion production and perception, e. g., [1], [2], [3] showing partially considerable language effects. Similarly, automatic recognition of emotion across languages has been approached showing the challenge of this task, cf., e. g., [4], [5], [6]. This is well in agreement with related speech processing task experiences across languages [7]. However, already in language acquisition of children, affect plays a crucial role as they prefer listening to happy speech, making it likely that affect generalises across language to some degree [8], which is also given evidence to in [9], [10], [11], [12]. In this contribution, we thus shed light on mis-matched vs matched condition emotion recognition across a variety of languages. To this end, let us first review the current state-of-play in availability of speech emotion databases and automatic recognition research focussing on language diversity.

Luckily, the number of databases dealing with voice characteristics such as emotion is increasingly covering various languages. Many databases have recently been developed in this field, making cross-lingual studies more and more feasible. However, many if not most of them are restricted, and only a few can be freely accessed. Today approximately 3 000 to 6 000 languages are spoken by humans. A group of languages that descends from a common ancestor is known as a language family. The most spoken languages in the world today belong to the Indo-European family (which includes languages such as English, Spanish, Russian); to the Sino-Tibetan languages, (which include Mandarin Chinese, Cantonese and many others), to Semitic languages (which include Arabic, Amharic, and Hebrew), and to Bantu languages (which include Swahili, Zulu, Shona, and hundreds of other languages spoken throughout Africa).

The purposes of the databases can be very broad. The emotions can reflect real differences in their vocal expression from speaker to speaker, from culture to culture [29], and across genders and situations. Depending of the goals of the database, many factors vary such as the number of speakers, the spoken language, the type of dialect, the gender of speakers, and the types of emotional states. Some features may be consistent across studies, others may be quite variable. The easiest way to collect emotional speech with known labels is to have actors which can simulate it. In fact, good actors can generate emotional speech such that listeners classify it with high agreement. For example, acted material studied in [59],

The research leading to these results has received funding from the European Union's Horizon 2020 Programme through Grants Nos. 645378, 644632, and 645094 (ARIA-VALUSPA, MixedEmotions, and SEWA).

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

TABLE II
OVERVIEW OF THE SELECTED DATABASES IN THE EXPERIMENTS (F/M: (FE-)MALE SUBJECTS).

Database	Language	Family	Symbol	# Arousal		# Valence		# m	# f	kHz
			-	+	-	+				
Emo-DB [32]	German	Germanic	DE	248	246	352	142	5	5	20
DES [61]	Danish	Germanic	DK	104	156	156	104	2	2	20
Enterface [20]	English	Germanic	GB	215	857	427	645	34	8	16
SES [62]	Spanish	Romanic	ES	15	18	15	18	1	0	16
SRoL [18]	Romanian	Romanic	RO	154	154	154	154	11	8	22
Busim [45]	Turkish	Turkic	TR	242	242	242	242	3	8	16
Mandarin [13]	Mandarin	Sino-Tibetan	CN	60	180	120	120	3	3	22
Burmese [13]	Burmese	Sino-Tibetan	MM	69	177	108	138	3	3	22

TABLE I
OVERVIEW OF THE MOST SPOKEN LANGUAGES BY PERCENTAGE OF NATIVE SPEAKERS (NS) IN THE WORLD AND ACCORDING RANK (SOURCE: NATIONALENCYKLOPEDIN 2010) WITH REFERENCES TO EXISTING EMOTIONAL SPEECH DATABASES IN THESE. ONLY SUCH WHERE ANNOTATED AND VALIDATED EMOTIONAL SPEECH DATA AND PARTIALLY RECOGNITION RESULTS ARE FOUND IN THE LITERATURE ARE CONTAINED. OVERALL, THIS COVERS FOR 66 % OF THE WORLD'S NATIVE LANGUAGE SPEAKING POPULATION.

Language	% NS	Rank	Reference
Mandarin	14.40	1	[13], [14], [15]
Spanish	6.15	2	[16], [17], [18], [19]
English	5.43	3	[17], [20], [21]
Hindi	4.70	4	[22], [23], [24]
Arabic	4.43	5	[25]
Portuguese	3.27	6	[26], [27]
Bengali	3.11	7	[24]
Russian	2.33	8	[28]
Japanese	1.90	9	[29]
Punjabi	1.44	10	[30], [23], [31]
German	1.39	11	[32], [33], [34]
Malay/Indonesian	1.16	14	[35]
Telugu	1.15	15	[36]
Vietnamese	1.14	16	[37]
Korean	1.14	17	[38]
French	1.12	18	[17], [39]
Marathi	1.10	19	[40], [41]
Tamil	1.06	20	[42]
Urdu	0.99	21	[43], [30], [31]
Persian	0.99	22	[44]
Turkish	0.95	23	[45]
Italian	0.90	24	[46]
Cantonese	0.89	25	[47]
Thai	0.85	26	[48]
Gujarati	0.74	27	[23]
Polish	0.61	30	[49]
Pashto	0.58	31	[31]
Burmese	0.50	38	[13]
Sindhi	0.39	47	[31]
Romanian	0.37	50	[50], [51]
Dutch	0.32	57	[52]
Assamese	0.23	67	[53], [54]
Hungarian	0.19	73	[55]
Greek	0.18	75	[56]
Czech	0.15	83	[57]
Swedish	0.13	91	[58], [21]
Balochi	0.11	99	[31]

produced human recognition rates of 78 % for hot anger, 76 % for boredom, and 75 % for interest, though scores for other emotions were lower with an average recognition rate of 48 % across 14 emotions. Clearly, however, there are differences between acted and non-acted emotional speech [60], which is why one wishes for the latter if the use-case is in an every-day-usage environment. Unfortunately, such non-acted emotions are less predictable and they can be difficult to collect in large sample volumes of various subjects with a specific emotional state. Inducing or enacting emotions is thus an often chosen avenue. The ideal case might be naturalistic emotional behaviour from real-life situations. However, such data are mostly private, and data found on the Internet, radio, and television on the other hand may is often copyright-protected.

Let us now give an overview on which languages covered in the research on data and recognition of emotion and speech. Table I gives a rough overview on languages that have been explored in automatic speech emotion recognition or where (validated) data is available – it also shows the percentage of the world's native speakers covered by the language and its rank in terms of this percentage of native speakers in the world. Beyond the languages shown picked from the list of the 100 languages with the highest number of native speakers according to the Swedish Nationalencyklopedin 2010, some further languages are found in studies dealing with computational analysis of emotional speech such as Danish [61], Finnish [63], Hebrew [64], [21] or Slovenian [17]. Besides, some databases exist that by intention have pseudo-language character – the most prominent example likely being the GEMEP corpus [65] that was featured in the Interspeech 2013 Computational Paralinguistics Challenge [66]. This makes it evident that almost half of the world's population is not yet covered lending hope to the usability of closely related languages to cover up for others. Beyond this overview in numbers, let us give some examples on characteristics of some representative emotional speech database next again emphasising on language diversity in order to provide a better impression on the variability of protocols followed and emotion categories contained: The emotional speech database in Japanese described in [29] consists in vowel consonant vowel (VCV) segments for each of the three emotions anger, sadness, and joy. These segments can generate any accent pattern of Japanese. The VCVs were collected from a corpus of 400 linguistic unbiased utterances. The utterances were analysed to derive a guideline for designing VCV databases, and to derive an equation for each phoneme, which can predict its duration

based on its surrounding phonemic and linguistic context. Twelve people judged the database and they recognised each emotion with a rate of 84 %. The Swedish emotional speech database featured in [58] contains speech in 9 emotion categories: joy, surprise, sadness, fear, shyness, anger, dominance, disgust, and neutral. Different nationality listeners classified the emotional utterances to an emotional state. The listener group consisted of 35 native Swedish speakers, 23 native Spanish speakers, 23 native Finnish speakers, and 12 native English speakers. The non-Swedish listeners were Swedish immigrants and all had knowledge of Swedish, of varying proficiency. An emotional speech corpus in Hebrew studied the following emotions: anger, fear, joy, sadness or disgust from a group of 40 students (19 males and 21 females). The speakers recalled an emotional event and tried to experience the same feelings as in the original event. It was measured also three physiological variables: the electromyogram,the heart rate and galvanic skin resistance. The goal was to determine a set of criteria that could represent each emotion [64]. The Russian affective language database consists of 10 sentences with different syntactic, structural and discourses types, which were read by 61 (12 male and 49 female) persons, aged between 16 and 28 years who are native speakers of Russian. The recordings were made following six affective-emotional states: neutral/unemotional, surprise, happiness, anger, sadness, and fear [28]. All the data were recorded on a portable Digital Audio Tape-recorder. The database serves as a source for developing and training a system of emotions recognition in Russian and provides data for designing a new system of Russian intonation description. The Italian database of emotional speech described in [46] includes isolated emotions and 'combined emotions'. The first part contains a set of Italian non-sense words, acted in the 'big six' emotional states – anger, disgust, fear, happiness, sadness, surprise, and added neutral, with three different intensity levels (low, medium, high). The second part includes significant examples of transitions from an emotional state to another during speech. A further part contains long sentences with a good coverage of Italian phonemes. The Interface databases is a multilingual collection of emotional speech. The aim of this database is to study the emotional speech as well as to analyse the emotion characteristics for speech synthesis and for automatic emotion classification purposes. They studied the big six emotions. The neutral tone was defined as a reference to emotional speech. The recordings were made by actors.The databases consist of 175–190 sentences for each language. The recordings have been performed in silent rooms using a high quality condenser microphones. The English Interface database contains 8 928 sentences, Slovenian 6 080 sentences, French 5 600, and Spanish 5 520 sentences [17]. These examples highlight the difficulties one is faced with when trying to research language effects in automatic speech emotion recognition: The corpora come with varying emotion categories and models, different number of speakers and samples, different chunking in time, different acoustic conditions, different degrees of naturalism, different spoken content variability reaching from prompted to free speech, to name but a few co-influencing factors that will be hard if not impossible to rule out entirely in cross-language analysis.

In the next section we present a number of further speech databases which are freely accessible on the Internet (some of them with an end user license agreement) and were thus selected in the experiments we describe later. They give results of our analyses regarding cross-language emotion recognition. The final section of this contribution then includes brief discussions and conclusions.

II. Speech Databases

Eight languages are covered in the databases described next that were selected for computational experiments on cross-language emotion recognition in the ongoing. Given the above described high variability in databases, these were chosen to be I) including clean speech and II) rather prototypical given the challenge of cross-language emotion recognition. Further selection criteria of these are availability, good overlap in contained emotion categories, and coverage of different (partially overlapping) language families. Obviously, one would wish for much more languages, more equal conditions, and other factors, but the sheer availability is the bottleneck in the young discipline of cross-language emotion recognition. As these sets come in different emotions, a mapping between categories is needed and is fulfilled here by binary arousal and binary valence mapping per emotion category. The chosen mapping is not unique but chosen in an intuitive manner. The chosen mapping is shown below for each database as follows: "emotion catgory (+/- Arousal / +/- Valence, # instances)". This mapping procedure was first suggested in [67] and has been repeatedly followed since when it comes to cross-corpus emotion analyses.

A. German Language

The Berlin Emotional Speech Database (Emo-DB) [32] database contains about 900 utterances spoken in seven emotions by 10 different actors. There are the sound files itself, the label files (syllable label files and phone label files), information about the results of different perception tests (including the recognition of emotions, the evaluation of naturalness, the syllable stress and the strength of the displayed emotions) as well as some results of the measurements of fundamental frequency, energy, loudness, duration, stress and rhythm available in the distribution. The emotions and speech samples usually chosen in studies (according to a validation study [32]) are: anger (+/-,127), boredom (-/-,79), disgust (-/-,38), fear (+/-,55), happiness (+/+,64), sadness (-/-,53), and neutral (-/+,78).

B. Danish Language

The Danish Emotional Speech (DES) database contains recordings from 4 actors (2 male and 2 female) expressing 5 emotions, each for 30 sec, thus totalling 10 min of Danish emotional speech. The data was recorded in an acoustically

TABLE III
SET OF 31 LLD AND 42 FUNCTIONALS. [1]NOT APPLIED TO DELTA LLD. [2]FOR DELTA LLD THE MEAN OF ONLY POSITIVE VALUES IS APPLIED, OTHERWISE THE ARITHMETIC MEAN IS APPLIED. [3]NOT APPLIED TO VOICING RELATED LLD.

Energy & spectral low-level descriptors (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, MFCC 1–10
Voicing related low-level descriptors (6)
F_0 (sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: 'jitter of jitter'), logarithmic Harmonics-to-Noise Ratio (logHNR)
Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 %–99 %, percentage of frames contour is above: minimum + 25%, 50%, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[3], standard deviation of segment length[3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
Linear Prediction (LP) gain, LP Coefficients 1 – 5

damped sound studio at Aarhus theatre. A high quality microphone was used, which did not influence the spectral amplitude or phase characteristics of the speech signal. Between the operator room and the recording room, a window was placed so that the actors and the operators could see each other at all times. The following was recorded: 2 single words, 9 sentences and 2 passages of fluent speech. The target voices should also record: 8 passages, 10 sentences spoken with a neutral voice [61]. The emotions and instances after typical chunking in the database are: angry (+/-,52), happy (+/+,52), sad (-/-,52), surprise (+/+,52), and neutral (-/+,52).

C. English Language

The Enterface database [20] contains recordings from 42 persons coming from 14 different nationalities (e. g., Belgium, Turkey, France, Spain, Greece, Italy, and Slovakia), with a percentage of 81 % male and 19 % female speakers. Each subject was told to listen to six successive short stories, each of them eliciting a particular emotion. They had to react to each of the situations. The indication given to the subject was to be as emotional as possible. The emotions and usually chosen instances contained in the database are: anger (+/-,215), fear (+/-,215), happiness (+/+,212), sadness (-/-,215), surprise (+/+,215).

D. Spanish Language

The Spanish Emotional Speech (SES) database [62] contains three sets of emotional recording sessions and two neutral sessions; each session includes three paragraphs, fifteen short sentences, and thirty isolated words, which have been read by a professional Spanish actor, simulating four emotions; the short sentences of the first set of recording sessions (one of each emotion and the neutral style 'one') have been manually pitch-marked and phonetically-labelled; further, the first two paragraphs of the first set of sessions have been manually pitch-marked and phonetically-labelled, except for anger [18]. The emotions and instances from this database are: angry (+/-,9), happy (+/+,9), sad (-/-,9), and neutral (-/+,6).

E. Romanian Language

The Spoken Romanian Language (SRoL) database [18] includes more than 1 000 recordings of spoken language, in different encoding formats and accompanied by annotations and extensive documentation. The database contains files with vowels, consonants, diphthongs, sentences with emotional states, linguistic particularities for the Romanian language, dialectal voices, and gnathosonic, and gnatophonic sounds. The registered sentences are: *mother is coming* (vine mama, in Romanian), *who did that?* (cine a facut asta, in Romanian), last night (Aseara, in Romanian), and *you came to me again* (yi venit iar la mine, in Romanian). The recordings were performed at a sampling frequency of 22 kHz with PCM signed (24 bits mono). The database contains also a recording technical protocol regarding information about the noise, the microphone used, the soundboard, and the corresponded drivers. The recordings are accompanied by the speaker profile and by a questionnaire concerning vocal pathology and objective factors for every speaker. The speakers are aged between 25–35 years; they are from the middle area of Moldova and have no manifested pathologies [50]. The emotions and instances in the database are: anger (+/-,77), joy (+/+,77), sadness (-/-,77), and neutral (-/+,77).

F. Turkish Language

The BUSIM SPG Turkish Emotional Database [45] contains 484 utterances (121 utterances per emotional state).The recordings were made by 11 different speakers (8 females, 3 males) that recorded 11 different Turkish sentences, and each sentence was recorded four times. Each utterance was recorded at 16 kHz, 16 bits and 256 kbps [45]. The emotional states and instances recorded are: anger (+/-,121), joy (+/+,121), sadness (-/-,121), and neutral (-/+,121).

G. Burmese and Mandarin Language

This database includes short utterances covering the six archetypal emotions. A total of six native Burmese language speakers and, six native Mandarin language speakers (3 females, 3 males, each) spoke 720 emotional utterances [13]. The speakers were recruited from university staff, postgraduate, and undergraduate students from two universities. Recording was executed in a laboratory room that was noise free. The

TABLE IV
UNWEIGHTED ACCURACY (UA) FOR CROSS-LANGUAGE POLARITY RECOGNITION; TRAIN ON ONE LANGUAGE, TEST ON ANOTHER LANGUAGE; MAIN DIAGONAL (*): INTRA-CORPUS CROSS-VALIDATION (NOT INCLUDED IN THE MEANS); ¬ INDICATES DATA-BASED MODEL-INVERSION CASES.

% UA	**AROUSAL** (train on:)								Mean	
test on:	DE	DK	GB	ES	RO	TR	CN	MM	UA¬	UA
DE	97.3*	50.3	71.3¬	54.6¬	50.0	69.0	60.9¬	68.7¬	60.7	44.8
DK	50.7¬	95.0*	79.4	60.6	59.0	50.0	78.3	85.7	66.2	66.0
GB	56.4	62.6	87.7*	63.6	59.7	50.2	75.4	71.1	62.7	62.7
ES	52.6	65.3	54.2	100.0*	53.2	51.8	77.0	76.4	61.5	61.5
RO	63.1	68.0	78.9	60.7¬	87.3*	52.8	65.4	54.0	63.3	60.2
TR	51.4	53.0	78.4	54.6¬	56.8	88.4*	72.9	50.8	59.7	58.4
CN	72.0	65.0	76.8	72.7	68.1	63.8	99.5*	92.6	73.0	73.0
MM	58.1¬	57.4¬	57.9	54.6¬	51.9	53.0	85.4	97.1*	59.8	54.0
UA¬	57.8	60.2	71.0	60.2	57.0	55.8	73.6	71.3	63.4	60.1
UA	55.2	58.1	64.9	53.2	57.0	55.8	70.5	66.0	60.1	
	VALENCE									
DE	86.3*	58.5¬	59.9	54.5	50.3	51.6	62.5	54.4	56.0	53.5
DK	50.8	68.4*	59.6¬	51.5	52.5	58.6	55.5¬	57.8¬	55.2	48.6
GB	71.0	53.9¬	79.4*	51.5	50.9	51.2	54.6¬	52.4	55.1	52.6
ES	58.3¬	54.2	61.3	100.0*	50.0	51.6	57.1¬	64.3¬	56.7	48.2
RO	61.3	52.0¬	57.0¬	54.5	56.4*	54.2¬	55.0¬	54.0	55.4	50.2
TR	67.2	57.3	50.4	51.6¬	52.9	72.3*	50.5¬	52.9¬	54.7	53.3
CN	57.7¬	54.6	54.2	54.5	50.7¬	54.9	95.8*	83.7	58.6	56.2
MM	51.3¬	51.6¬	51.7	54.5	53.6¬	50.7¬	77.0	94.7*	55.8	53.7
Mean UA¬	59.7	54.6	56.3	53.2	51.6	53.3	58.9	59.9	55.9	52.1
Mean UA	54.7	50.0	51.6	52.8	50.3	51.9	52.4	52.8	52.1	

TABLE V
MEAN UNWEIGHTED ACCURACY (UA) WITHIN THE SAME LANGUAGE (L), AND WITHIN/ACROSS LANGUAGE FAMILY (LF).

% UA	same L	within LF	across LF
Arousal	94.0	66.3	62.7
Valence	81.7	61.9	54.6

speakers were left alone throughout the recording session. All speech data are coded at 16 bit/sample and sampled at 22 kHz. The emotions and instances from the Burmese database are: fury (+/-,69), joy(+/+,69), surprise (+/+,69), and sadness (-/-,39), and from Mandarin database: fury (+/-,60), joy (+/+,60), surprise (+/+,60), and sadness (-/-,60).

The emotional speech databases analysed in this study are summarised in Table II.

III. EXPERIMENTAL RESULTS

In this section, we want to demonstrate some tendencies of cross-language emotion reocgnition, as exemplified by the eight databases described in section II. Overall, we train and test each database against each, resulting in 56 tuples, plus 8 intra-database runs in 10-fold cross-validation. For these experiments, we employ a well-standardised acoustic feature vector: The set used is our openSMILE toolkit's (version 1.0.1) AVEC set [68] with 1 941 features brute forced by functional application to low-level descriptors (LLD). Details for the LLD and functionals are given in Table III. The set of LLD covers a standard range of commonly used features in speech emotion recognition. The approach is based on brute-forcing by calculating LLD, adding their deltas coefficients, yet avoiding LLD/functional combinations that produce values which are constant, contain (very) little information, and/or high amount of noise (cf. [68] for details). Features are computed per whole speech clip. As machine learning algorithm we employ one-vs-one class support vector machines (SVM) trained by sequential minimal optimisation with linear kernel and a complexity parameter of 0.5 using the WEKA 3 implementation [69]. The rationale behind these choices for the feature extraction and classification is highest reproducibility and standardisation, as these choices accompany the Interspeech and AVEC series of challenges on emotion recognition, cf., e. g., [51], [68]. Accordingly, no further optimisation is carried out to provide a transparent and redoable experiment rather than 'tweaking and tuning' to 'quench out' some percentage points in accuracy. However, we found that the models are partially translating poorly across languages leading to considerably sub-chance level accuracies. This made it necessary to apply a simple rule-based inversion of the (binary) target classes for arousal and valence as follows: Based on 10 % of the target data, a decision is made whether or not to swap classes from the learnt model. In the results shown in Table IV for binary arousal and valence classification, these decisions are highlighted by ¬. In addition, overall mean results are given with and without this strategy. In these tables, the numbers outside the main diagonal represent the (mis-matched) cross-language tests, i. e., one corpus is used as test set and another is used for training, each. On the main diagonal, results for within corpus classification based on cross-validation is given – obviously only as a reference. As a measure of comparison, we use unweighted accuracy (UA), i. e., the recall per (each of the two) class divided by the number of classes (here simply two). This procedure has

become popular in emotion recognition, as it well takes the usual imbalance across classes into account. Just as the feature extractor and classifier implementations, it has been used in various challenges in the field [51], [68].

IV. Discussion and Conclusion

Clearly, the results presented in Table IV have to be taken with a grain of salt and interpreted with utmost care. They shall mostly serve as tendencies, given the limitations described above in more detail that one is faced with due to availability of multilingual emotional speech data these days. Comparing the values for UA and UA$^\neg$ one sees an absolute delta of 3.3 (arousal) and 3.8 (valence) percent points for processing with and without additional post-processing of the learnt SVM models by rule-based model inversion based on a 10 % sample of the data. This shows that on average, this is an efficient step when dealing with cross-language emotion recognition. Further, one can group the results by language families as indicated by the grids in Table IV. Average results for within and across language family recognition are shown in Table V. The absolute delta between within and across language family is 3.6 (arousal) and 7.3 (valence) percent points UA. Comparing this to the overall mean recognition rate of arousal vs valence it shows that not only is arousal easier to recognise from acoustics – a well-known fact in the field (cf., e. g., [51]) – but also it seems that valence generalises less across language families. The additional summary of within language results should not be directly compared with these numbers, as it is not coming from a cross-corpus setting – rather, it serves to demonstrate again the easier recognition of arousal rather than valence. One also finds some interesting details in the result tables such as highly encouraging pairs of languages across language family, such as when training arousal on Burmese speech and testing on Danish leading to the best cross-language family constellation in the table. Likewise, we conclude that cross-language and even cross-language family acoustic emotion recognition is feasible, but it will remain best to have a suited language resource at hand for each desired target language.

Obviously, one needs to redo similar experiments with more languages under more equal conditions given data availability. For future work, we further consider transfer learning across languages of particular interest, as has recently been shown successful to adapt adult emotional speech data to children's speech [70] or even to train a speech emotion classifier with music [71].

References

[1] J. J. Ohala *et al.*, "An ethological perspective on common cross-language utilization of f0 of voice," *Phonetica*, vol. 41, no. 1, pp. 1–16, 1984.

[2] A. Tickle, "English and japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[3] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[4] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *International Journal of Speech Technology*, vol. 6, no. 3, pp. 311–320, 2003.

[5] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection," *Speech-Prosody, Chicago, USA*, 2010.

[6] M. Bhaykar, J. Yadav, and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm," in *Communications (NCC), 2013 National Conference on*. IEEE, 2013, pp. 1–5.

[7] P. Fung and T. Schultz, "Multilingual spoken language processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 89–97, 2008.

[8] L. Singh, J. L. Morgan, and K. S. White, "Preference and processing: The role of speech affect in early spoken word recognition," *Journal of Memory and Language*, vol. 51, no. 2, pp. 173–189, 2004.

[9] 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.

[10] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in cantonese and english," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1394–1405, 2009.

[11] 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.

[12] H. S. Cheang and M. D. Pell, "Recognizing sarcasm without language: A cross-linguistic study of english and cantonese," *Pragmatics & Cognition*, vol. 19, no. 2, pp. 203–223, 2011.

[13] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[14] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and P.-J. Li, "Mandarin emotional speech recognition based on svm and nn," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1096–1100.

[15] 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.

[16] J. Montero, J. Gutiérrez-Arriola, J. Colás, E. Enríquez, and J. Pardo, "Analysis and modelling of emotional speech in spanish," in *ICPhS*, vol. 2, 1999, pp. 957–960.

[17] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: Design and collection of a multilingual emotional speech database." in *LREC*, 2002.

[18] S. Feraru, H. Teodorescu, and M. Zbancioc, "Srol-web-based resources for languages and language technology e-learning," *International Journal of Computers Communications & Control*, vol. 5, no. 3, pp. 301–313, 2010.

[19] V. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 0038–45, 2013.

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

[21] E. Marchi, B. Schuller, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, O. Golan, S. Friedenson, S. Tal, S. Bölte, S. Berggren, D. Lundqvist, and M. S. Elfström, "Voice Emotion Games: Language and Emotion in the Voice of Children with Autism Spectrum Condition," in *IDGEI 2015 as part of IUI 2015*. Atlanta, GA: ACM, 2015, 9 pages.

[22] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "Iitkgp-sehsc: Hindi speech corpus for emotion analysis," in *Devices and Communications (ICDeCom), 2011 International Conference on*. IEEE, 2011, pp. 1–5.

[23] S. Agrawal, "Emotions in hindi speech-analysis, perception and recognition," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*. IEEE, 2011, pp. 7–13.

[24] K. S. Rao and S. G. Koolagudi, "Identification of hindi dialects and emotions using spectral and prosodic features of speech," *IJSCI: International Journal of Systemics, Cybernetics and Informatics*, vol. 9, no. 4, pp. 24–33, 2011.

[25] W. M. Azmy, S. Abdou, and M. Shoman, "Arabic unit selection emotional speech synthesis using blending data approach," *International Journal of Computer Applications*, vol. 81, no. 8, pp. 22–28, 2013.

[26] S. L. Castro and C. F. Lima, "Recognizing emotions in spoken language: A validated set of portuguese sentences and pseudosentences for research on emotional prosody," *Behavior Research Methods*, vol. 42, no. 1, pp. 74–81, 2010.

[27] C. F. Lima and S. L. Castro, "Speaking to the trained ear: musical expertise enhances the recognition of emotions in speech prosody." *Emotion*, vol. 11, no. 5, p. 1021, 2011.

[28] V. Makarova and V. A. Petrushin, "Phonetics of emotion in russian speech," in *XVth international conference of phonetic sciences*, 2003.

[29] Y. Niimi, M. Kasamatsu, T. Nishinoto, and M. Araki, "Synthesis of emotional speech using prosodically balanced vcv segments," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[30] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2. IEEE, 2005, pp. ii–1125.

[31] S. A. Ali, S. Zehra, M. Khan, and F. Wahab, "Development and analysis of speech emotion corpus using prosodic features for cross linguistics," *International Journal of Scientific & Engineering Research*, vol. 4, no. 1, 2013.

[32] 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.

[33] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[34] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus," in *LREC*, 2008, pp. 28–31.

[35] A. A. Razak, M. H. M. Yusof, and R. Komiya, "Towards automatic recognition of emotion in speech," in *ISSPIT*. IEEE, 2003, pp. 548–551.

[36] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "Iitkgp-sesc: speech database for emotion analysis," in *Contemporary Computing*. Springer, 2009, pp. 485–492.

[37] T. D. Ngo and T. D. Bui, "A study on prosody of vietnamese emotional speech," in *Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on*. IEEE, 2012, pp. 151–155.

[38] S.-J. Kim, K.-K. Kim, H. B. Han, and M. Hahn, "Study on emotional speech features in korean with its application to voice conversion," in *Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 342–349.

[39] B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers, "Cinemo-a french spoken language resource for complex emotions: Facts and baselines." in *LREC*, 2010.

[40] V. N. Degaonkar and S. D. Apte, "Emotion modeling from speech signal based on wavelet packet transform," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 1–5, 2013.

[41] V. B. Waghmare, R. R. Deshmukh, P. P. Shrishrimal, and G. B. Janvale, "Emotion recognition system from artificial marathi speech using mfcc and lda techniques," in *Fifth International Conference on Advances in Communication, Network, and Computing–CNC*, 2014.

[42] J.-S. Park, J.-H. Kim, and Y.-H. Oh, "Feature vector classification based speech emotion recognition for service robots," *Consumer Electronics, IEEE Transactions on*, vol. 55, no. 3, pp. 1590–1596, 2009.

[43] Y. Wang and L. Guan, "An investigation of speech-based human emotion recognition," in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*. IEEE, 2004, pp. 15–18.

[44] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications*, vol. 49, no. 2, pp. 277–297, 2010.

[45] H. Ekenel, H. Meral, and A. Ozsoy, "Analysis of emotion in turkish," in *XVII. National Conference on Turkish Linguistics*, 2003.

[46] N. Mana, P. Cosi, G. Tisato, F. Cavicchio, E. C. Magno, and F. Pianesi, "An italian database of emotional speech and facial expressions," in *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006*, 2006, p. 68.

[47] X. Zuo, L. Lin, and P. Fung, "A multilingual database of natural stress emotion," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 1174–1178.

[48] I. Stankovic, M. Karnjanadecha, and V. Delic, "Improvement of thai speech emotion recognition by using face feature analysis," in *ISPACS*. IEEE, 2011, pp. 1–5.

[49] P. Staroniewicz, "Recognition of Emotional State in Polish Speech-Comparison between Human and Automatic Efficiency," in *Biometric ID Management and Multimodal Communication*. Springer, 2009, pp. 33–40.

[50] H.-N. Teodorescu and S. M. Feraru, "A study on speech with manifest emotions," in *Text, Speech and Dialogue*. Springer, 2007, pp. 254–261.

[51] 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 Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[52] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *1st Richmedia Conference*, 2003, pp. 109–119.

[53] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from assamese speeches using mfcc features and gmm classifier," in *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE, 2008, pp. 1–5.

[54] ——, "Vocal emotion recognition in five native languages of assam using new wavelet features," *International Journal of Speech Technology*, vol. 12, no. 1, pp. 1–13, 2009.

[55] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer, 2008, pp. 213–224.

[56] T. Kostoulas, T. Ganchev, I. Mporas, and N. Fakotakis, "A Real-World Emotional Speech Corpus for Modern Greek," in *LREC*, 2008.

[57] M. Gruber and M. Legát, "Single speaker acoustic analysis of czech speech for purposes of emotional speech synthesis," in *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, vol. 2, 2008.

[58] Å. Abelin and J. Allwood, "Cross linguistic interpretation of emotional prosody," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[59] H. R. Markus and S. Kitayama, "The cultural construction of self and emotion: Implications for social behavior," *Emotions in social psychology: Essential reading*, pp. 119–137, 2001.

[60] S. Whiteside, "Acoustic characteristics of vocal emotions simulated by actors," *Perceptual and motor skills*, vol. 89, no. 3f, pp. 1195–1208, 1999.

[61] I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database des," *Internal AAU report, Center for Person Kommunikation, Denmark*, p. 22, 1996.

[62] J. M. Montero, J. M. Gutierrez-Arriola, S. E. Palazuelos, E. Enriquez, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: from speech database to tts." in *ICSLP*, vol. 98, 1998, pp. 923–926.

[63] J. Toivanen, T. Seppänen, and E. Väyrynen, "Automatic recognition of emotions in spoken finnish: preliminary results and applications," in *International AAI Workshop on Prosodic Interfaces*. Citeseer, 2003, pp. 85–89.

[64] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in hebrew based on objective criteria," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[65] T. Bänziger, H. Pirker, and K. Scherer, "Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions," in *LREC*, vol. 6, 2006, pp. 15–019.

[66] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Interspeech*, 2013.

[67] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *ASRU*. IEEE, 2009, pp. 552–557.

[68] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 415–424.

[69] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[70] 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.

[71] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE, 2014, pp. 3592–3598.