MixedEmotions

Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets

D4.1 Emotion Recognition from Multilingual Audio Content, initial version

Project ref. no	H2020 644632
Project acronym	MixedEmotions
Start date of project (dur.)	01 April 2015 (24 Months)
Document due Date	31 December 2015 (Month 9)
Responsible for deliverable	University of Passau
Reply to	bjoern.schuller@imperial.ac.uk
Document status	Final

Project reference no.	H2020 644632
Project working name	MixedEmotions
Project full name	Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets
Document name	MixedEmotions_D4.1_31_12_15_Emotion Recognition from Multilingual Audio Content initial version_UP
Security (distribution level)	PU
Contractual delivery date	31 December 2015
Deliverable number	D4.1
Deliverable name	Emotion Recognition from Multilingual Audio Content, initial version
Type	Other
Version	Final
WP / Task responsible	WP4 / University of Passau
Contributors	Hesam Sagha, Björn Schuller, Maryna Gavryukova
EC Project Officer	Susan Fraser

Table of Contents

Executive Summary

MixedEmotions aims to develop a platform able to recognize emotions from multilingual multi-modal (audio, video, and text) data. This deliverable (D4.1) focuses on the multilingualism aspect of the project through audio content. The content of this deliverable is twofold: first we investigate and compare cross-lingual emotion recognition for "the same language", "within language family", and "between language families." Second, we propose a transfer learning method to enhance emotion recognition between different languages (corpora) in order to re-use the knowledge gained for one language and apply it to another language. The implementation of the method in the MixedEmotions platform is awaiting final platform specification.

1 Introduction

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 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. In Section 2 we overview the languages covered in emotion and speech recognition research and investigate the performance of emotion recognition between the corpora in the same language, within same language family, and between different language family.

Moreover, in the MixedEmotions platform, since audio data could be in different languages, aiming to develop models for each language could be infeasible due to unavailability of ground truth labels as well as variety in the recording conditions (e. g., microphone, studio, home). In Section 3, we propose a transfer learning method to cope with this limitation. Therefore, we can re-use the information/model obtained in one source (which is annotated) for another source (which is *not* annotated). We have used '*canonical correlation analysis (CCA)*' for this goal. CCA tries to reduce dissimilarity between two distributions. To validate the method, pair-wise transfer learning between four emotional speech corpora with different languages (English, German, Italian, and Polish) is evaluated. We compare our approach with another state-of-the-art method (Shared-Hidden-Layer Auto-Encoder). On average the proposed method yields higher classification performance.

2 Cross-Language Acoustic Emotion Recognition: An Overview and Some Tendencies[1]

In this section 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 mismatched vs. matched condition

[1] The content of this section is from:
Feraru, Silvia Monica, Dagmar Schuller, and Björn Schuller. "Cross-Language Acoustic Emotion Recognition: An Overview and Some Tendencies." The sixth International Conference on Affective Computing and Intelligent Interaction (**ACII**), 2015

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.

2.1 Overview

A number of studies have investigated the effects of multiband cross-language human emotion production and perception, e. g., [1]–[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]–[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]–[12]. In this contribution, we thus shed light on mismatched 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 [13], 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 [14], 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

[15], 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 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 2-1 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 [16], Finnish [17], Hebrew [18], [19] or Slovenian [20]. Besides, some databases exist that by intention have pseudo-language character – the most prominent example likely being the GEMEP corpus [21] that was featured in the Interspeech 2013 Computational Paralinguistics Challenge [22]. 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 [13] 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

Table 2-1) Overview of the most spoken languages by percentage of native speakers (NS) in the world and according rand (source: NATIONALENCYKLOPEDIN 2010). These languages are covered in the literatures. This covers for 66% of the world's native language speaking population.

Language	%NS	Rank
Mandarin	14.40	1
Spanish	6.15	2
English	5.43	3
Hindi	4.7	4
Arabic	4.43	5
Portuguese	3.27	6
Bengali	3.11	7
Russian	2.33	8
Japanese	1.90	9
Punjabi	1.44	10
German	1.39	11
Malay/Indonesian	1.16	14
Telugu	1.15	15
Vietnamese	1.14	16
Korean	1.14	17
French	1.12	18
Marathi	1.10	19
Tami	1.06	20
Urdu	0.99	21
Persian	0.99	22
Turkish	0.95	23
Italian	0.90	24
Cantonese	0.89	25
Thai	0.85	26
Gujarati	0.74	27
Polish	0.61	30
Pashto	0.58	31
Burmese	0.50	38
Sindhi	0.39	47
Romanian	0.37	50
Dutch	0.32	57
Assamese	0.23	67
Hungarian	0.19	73
Greek	0.18	75
Czech	0.15	83
Swedish	0.13	91
Balochi	0.11	99

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 [23] 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 [18]. 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 [24]. 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 [25] 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 database 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 high quality condenser microphones. The English Interface database contains 8,928 sentences, Slovenian 6,080 sentences, French 5,600, and Spanish 5,520 sentences [20]. 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 section 2.2 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 section 2.4 of this contribution then includes brief discussions and conclusions.

2.2 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 category (+/- Arousal / +/- Valence, # instances)". This mapping procedure was first suggested in [26] and has been repeatedly followed since when it comes to cross-corpus emotion analyses.

2.2.1 German Language

The Berlin Emotional Speech Database (Emo-DB) [27] 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 [27]) are: anger (+/-,127), boredom (-/-,79), disgust (-/-,38), fear (+/-,55), happiness (+/+,64), sadness (-/-,53), and neutral (-/+,78).

2.2.2 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 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 [16]. The emotions and instances after typical chunking in the database are: angry (+/-,52), happy (+/+,52), sad (-/-,52), surprise (+/+,52), and neutral (-/+,52).

2.2.3 English Language

The Enterface database [28] 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).

2.2.4 Spanish Language

The Spanish Emotional Speech (SES) database [29] 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 [30]. The emotions and instances from this database are: angry (+/-,9), happy (+/+,9), sad (-/-,9), and neutral (-/+,6).

2.2.5 Romanian Language

The Spoken Romanian Language (SRoL) database [30] 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 (ai 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 [31]. The emotions and instances in the database are: anger (+/-,77), joy (+/+,77), sadness (-/-,77), and neutral (-/+,77).

Table 2-2) Overview of the selected databases in the experiments (F/M: (Fe-)male subjects)

Database	Language	Family	Symbol	#Arousal		#Valence		#m	#f	kHz
				+	-	+	-			
Emo-DB [27]	German	Germanic	DE	248	246	352	142	5	5	20
DES [16]	Danish	Germanic	DK	104	156	156	104	2	2	20
Enterface [28]	English	Germanic	GB	215	857	427	645	34	8	16
SES [29]	Spanish	Romanic	ES	15	18	15	18	1	0	16
SRoL [30]	Romanian	Romanic	RO	154	154	154	154	11	8	22
Busim [32]	Turkish	Turkic	TR	242	242	242	242	3	8	16
Mandarin [33]	Mandarin	Sino-Tibetan	CN	60	180	120	120	3	3	22
Burmese [33]	Burmese	Sino-Tibetan	MM	69	177	108	138	3	3	22

2.2.6 Turkish Language

The BUSIM SPG Turkish Emotional Database [32] 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 [32]. The emotional states and instances recorded are: anger (+/-,121), joy (+/+,121), sadness (-/-,121), and neutral (-/+,121).

2.2.7 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 [33]. 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 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 2-2.

2.3 Experimental results

In this section, we want to demonstrate some tendencies of cross-language emotion recognition, as exemplified by the eight databases described in section 2.2. 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 [34] with 1,941 features brute forced by functional application to low-level descriptors (LLD). Details for the LLD and functionals are given in Table 2-3. 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. [34] 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 [35]. 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., [36], [34].

Accordingly, no further optimisation is carried out to provide a transparent and re-doable 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 2-4 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.

Table 2-3) Set of 31 LLD and 42 functionals. [1]Not applied to delta LLD. [2]For delta LLD the mean of only possitive values are applied, otherwise the arithmatic 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)
F0 (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

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 [34], [36].

2.4 Discussion and future work

Table 2-4) 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); ¬ indicated data-based model-inversion cases.

% UA test on:	Arousal (train on:)								Mean	
	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	60.0	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*	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	61.1
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*	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¬	554.7	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	

Clearly, the results presented in Table 2-5 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¬ 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., [36]) – 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 (next section), as has recently been shown successful to adapt adult emotional speech data to children's speech [37] or even to train a speech emotion classifier with music [38].

Table 2-5) 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

3 Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace[2]

The fast pace of progress in the ubiquitous Internet facilitates collection of more data in less time. This is beneficial for the machine learning domain where more data can represent better the feature distribution. However, the side effect is that the labels of the collected data may not be available and they need to be annotated by human effort. This could be costly, tedious, cumbersome, and time consuming. Semi-supervised approaches such as active learning try to reduce this effort by automatic labelling of the data which have a high probability in a class and get the label from human when the certainty is not adequate. Nevertheless, for big databases this approach will be also less practical. Instead, transfer learning (TL) approaches try to use the knowledge which is already gained from other databases and use this knowledge for a new database [39]. Therefore, no more human effort would be needed for the annotation. In addition, in TL there is no necessity to hold the assumption of having the same distribution for training and testing data. This is beneficial when the train and test data are not obtained in the same way (e.g., studio vs. real environment audio recording) or their modalities are different (e.g., image vs. speech). TL approaches are categorized based on the domain and the task of the source and the target corpora. If both domain and task are the same and the target corpus is not annotated, the problem is called *domain adaptation* [40].
Kernel Mean Matching (KMM) was proposed to deal with domain adaptation problems by directly estimating the resampling weights by matching training and test distribution means in a reproducing kernel Hilbert space [41]. Recently, it was applied to reduce the acoustic and speaker difference across training and test data for speech emotion recognition [42]. Moreover, Deng et al. proposed the use of Shared-Hidden-Layer Auto-Encoders (SHLA) to obtain shared views of source and target emotional speech corpora [43]. In this approach, having two data corpora $\chi_{tr} \in \mathbb{R}^{n \times Q}$ and $\chi_{te} \in \mathbb{R}^{m \times Q}$, an artificial neural network with Q neurons in the input layer, H < Q neurons in the hidden layer, and 2Q neurons in the output layer is created (cf. Figure 3-1). A gradient descent approach is performed to find the tuning parameters. Finally, the outputs of the hidden

[2] The content of this section is from:
H. Sagha, J. Deng, M. Gavryukova, J. Han, Björn Schuller, "Cross-lingual speech emotion recognition using canonical correlation analysis on principal component subspace." The 41st IEEE international Conference on Acoustic, Speech and Signal Processing (ICASSP), 2016

layer are used for the training and classification. They compared this method with KMM, showing an improvement in cross-lingual emotion classification from speech.

In this paper, motivated by the success of SHLA, we propose a domain adaption method, which applies Canonical Correlation Analysis (CCA) to find the views with the highest correlations between source and target corpora. CCA is a statistical method to find linear bases so that the correlations between the projections of the variables onto these bases are mutually maximized [44]. Kernel CCA (KCCA) has been widely used for multimodal dimensionality reduction, such as for fMRI analysis [45] and speaker identification [46].

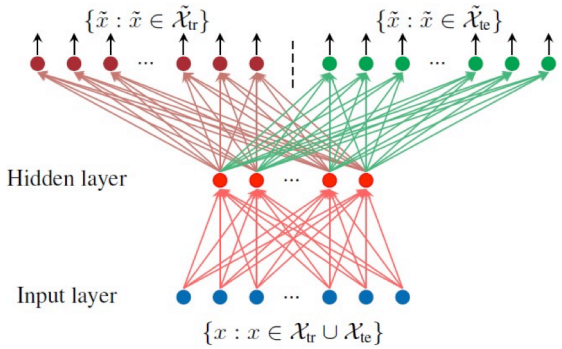

Figure 3-1) Domain adaptation using Shared-Hidden-Layer Auto-Encoder

Further, it was found that CCA based feature reduction can overcome the problem of over-fitting and provide a compact set of high quality features for computational paralinguistic applications [47]. Additionally, it has also been used as multi-view transfer learning for cross-language information retrieval where a parallel corpus is generated by text translation [48], [49].

Rather than using CCA as a feature reduction, we extend CCA to alleviate the mismatch between different languages for emotion recognition from speech. We generate two subspaces for each training and test corpora and map data onto each subspace. Therefore, two paired corpora are generated. Finally, we use Kernel CCA to find the views of the two subspaces where their mappings onto those views have the highest match.

The remainder of this deliverable is organized as follows. Next section provides the basis for CCA and Kernel CCA and our approach to apply it on domain adaptation. In Section 3.2 we describe the databases and in Section 3.3 we compare the results with SHLA. Section 5 draws conclusions and point future directions.

3.1 Method

Similar to Auto-Encoder transfer learning, the idea is to seek shared representation of features for the source and target databases. Then, a model is built on these features from the source database, and is used to label the target database. In the following, first we introduce general CCA, and the Kernel CCA, and then the proposed approach on how to deploy Kernel CCA for transfer learning.

3.1.1 CCA

Consider two databases $X \in R^{n \times d}$ and $Y \in R^{n \times p}$ having n paired multivariate random vector (x_i, y_i). CCA finds mappings (views) for X and Y so that the mapped data are highly correlated. In other words, CCA maximizes:

$$\rho = \max_{w,v} corr(Xw, Yv) = \max_{w,v} \frac{w' C_{xy} v}{\sqrt{w' C_{xx} w v' C_{yy} v}} \qquad (1)$$

Here, w can be found through Lagrangian approach and is the eigenvectors in the form of:

$$C_{xy}C_{yy}^{-1}C_{yx} = \lambda^2 C_{xx}w \quad (2)$$

where C_{xy} is the cross covariance matrix between X and Y. Then, we select the *N* vectors corresponding to the *N* largest eigenvalues. v is equal to

$$v = \frac{C_{yy}^{-1}C_{yx}w}{\lambda} \quad (3)$$

Xw and Yv have the highest correlation on the vector corresponding to the largest eigenvalue, and the second highest on the second vector corresponding to the second largest eigenvalue and so on. The upper limit of *N* is the minimum of rank of *X* and *Y*.

3.1.2 Kernel CCA (KCCA)

Kernel CCA, defines a Kernel on data and similar to CCA it seeks to maximize the correlation between mappings of these kernels:

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{\alpha' K_x^2 \alpha \beta' K_y^2 \beta'}} \quad (4)$$

where K_x and K_y are the kernel matrices corresponding to the two representations. As linear kernel they are $K_x = XX'$ and $K_y = YY'$, and as RBF kernels they are defined as $K_x(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$, where σ is a free parameter. The solution to (4) is in the form of eigenproblem:

$$(K_x + \kappa I)^{-1} K_y (K_y + \kappa I)^{-1} K_x \alpha = \lambda^2 \alpha \quad (5)$$

where κ is the regularization parameter. Moreover,

$$\beta = \frac{(K_y + \kappa I)^{-1} K_x \alpha}{\lambda} \quad (6)$$

Finally, the two mapping vectors are $w = X'\alpha$ and $v = Y'\beta$ [50].

3.1.3 KCCA-based domain adaptation

CCA and KCCA are useful when x_is and y_is are paired together. For example, Kaya et al. used them for feature reduction where x_i are the features and y_i are the binarized class labels [47]. Different from the use of KCCA for feature reduction, this paper makes use of KCCA in conjunction with Principal Component Analysis (PCA) for domain adaptation. Note that PCA is used to create two representations of each corpus on two sets of orthogonal vectors, as principal components, to preserve information on lower dimensions. The schematic of the approach is shown in Figure 3-2. First, source data is mapped on its principal components, X^{p_x}, and on target principal components, X^{p_y}. Similarly, target data is mapped on its principal components, Y^{p_y}, and on source principal components, Y^{p_x}. Then we reduce dimensions for each mapping to keep 99% of variation on principal components (This will help to avoid singularity during KCCA process). In general, there is no need to have the two views with the same number of transferred features. However, for the following analyses we kept them the same, equal to the maximum of the two reduced dimensions. Secondly, we find the shared view between the paired mapped data on the source principal components, $[X^{p_x}; Y^{p_x}]$ and

the target principal components $[X^{p_y}; Y^{p_y}]$ using canonical correlation analysis. Then, we pick top *N* dimensions (with largest correlations) of w and v and map X^{p_x} on w and Y^{p_y} on v. Finally, a classifier is trained on the mapped training data and tested on the mapped test data.

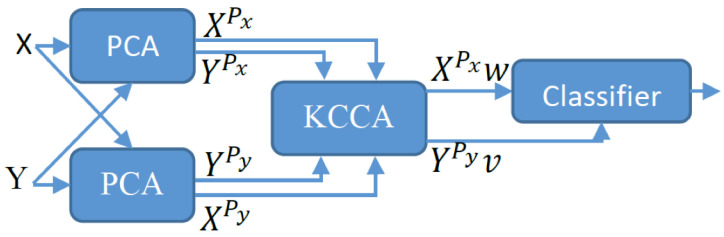

Figure 3-2) proposed approach. Superscript P_x and P_y denote the data mapping to Xs and Ys principal components, respectively.

Note that, if we do not map the data on their principal components or we use the same mapping for both corpora (i.e., $[X^{p_x}; Y^{p_x}] = [X^{p_y}; Y^{p_y}]$), then KCCA will be reduced to Kernel PCA.

3.2 Experiment

We compare the results of using no domain adaptation with five models of transfer learning on four emotional speech databases. Two of the domain adaptation methods are the proposed KCCA approach with RBF and linear kernels. We set N = 30, 40, 50 to create three models and we used Bayesian fusion to combine the decisions of each model. We tried κ value with 10, 50, and 100. The two other approaches are based on Kernel PCA with linear and RBF kernels. In this case we map data on three subspaces; X's principal components, Y's principal components, and $[X; Y]$'s principal components. We perform the classification on these three views and combine the decisions. Additionally, we compare the methods with the SHLA with the same number of hidden layers (N = 30, 40, 50) and decision fusion approach. We ran this process ten times and provided the average of the results.

3.2.1 Databases

Four databases with different languages have been investigated. Some information about these databases are provided in Table 3-1. All utterances of all corpora are generated by actors/actresses in studio environment. EMODB is a German emotional speech corpus where 10 sentences with emotionally neutral content is uttered in different emotions. In the SAVEE corpus each actor uttered 15 sentences in different emotions and they are validated by 10 subjects. The Italian corpus (EMOVO) contains utterances of 14 sentences simulating six emotional states plus neutral state. In the Polish Emotional Speech Dataset each speaker uttered five different sentences with six types of emotional load.

Table 3-1) Corpora information and the mapping of class labels into Negative/Positive valence. (#m): number of male speaker, (#f): number of female speakers, (Rate): Sampling rate.

Corpus	Language	#m	#f	Rate	Negative Valence (#)		Positive Valence (#)	
EMODB [27]	German	5	5	16	Anger, Sadness, Fear, Disgust, Boredom	(385)	Neutral, Happiness	(150)
SAVEE [51]	English	4	0	44	Anger, Sadness, Fear, Disgust	(240)	Neutral, Happiness, Surprise	(240)
EMOVO [52]	Italian	3	3	44	Anger, Sadness, Fear, Disgust	(336)	Neutral, Joy, Surprise	(252)

Polish [53]	Polish	4	4	44	Anger, Sadness, Fear, Boredom	(160)	Neutral, Joy	(80)

3.2.2 Feature Extraction

We extracted 384 features as in InterSpeech Emotion Challenge 2009 using OpenSMILE [54]. It comprises 12 functional of 2x16 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 coefficient 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 removed the features which are highly correlated with each other ($|\rho| > 0.95$) or if they have small variance ($< 10^{-10}$). This feature pruning keeps 311 features. Moreover, we removed the outlier data where a feature value is larger than 10 times of the standard deviation. Finally, we performed subject based normalization followed by corpus-based normalization (SC-normalization) which is shown to boost cross-language emotion recognition [55]. The SC-normalized data is fed to the transfer learning method.

3.3 Result

Table 3-2 shows the performance of classification using Simple Logistic classifier. The choice of this non-parametric classifier was to avoid parameter tuning and have a fair comparison between databases. Unweighted Average Recall (UAR) is used as performance measure. Classifications without transfer learning are denoted as '*Direct C*' for corpus normalized data and '*Direct SC*' for SC-normalized data. Only in one case (out of 12) the performance has not been improved by transfer learning. In 7 cases (out of 11) KCCA, in 3 cases SHLA and in 1 case PCA provide the highest accuracy. On average *Direct SC* yields 2.5% improvement with respect to *Direct C*. Furthermore, KCCA (Linear), KCCA (RBF), KPCA (Linear), KPCA (RBF) and SHLA yield 2.89%, 2.76%, 2.74%, 2.57%, and 1.98% average improvement over Direct SC, respectively.

The advantage of the KCCA and KPCA over SHLA is the analytical solution instead of gradient descent. Therefore, there is no risk of falling in a local minima and the learning speed is faster.

Table 3-2) UAR of transfer learning methods

		EMODB	SAVEE	EMOVO	Polish
EMODB	Direct C		55.83	56.40	71.87
	Direct SC		59.17	**58.73**	69.06
	KCCA (Lin.)		64.58	57.89	**72.81**
	KCCA (RBF)		**65.21**	56.99	72.50
	KPCA (Lin.)		64.37	57.64	72.19
	KPCA (RBF)		63.54	57.19	70.00
	SHLA		63.69	56.79	59.72
SAVEE	Direct C	62.46		54.56	65.00
	Direct SC	63.14		57.59	62.81
	KCCA (Lin.)	70.57		59.32	69.37
	KCCA (RBF)	**71.89**		**59.52**	67.50
	KPCA (Lin.)	70.09		58.73	72.81
	KPCA (RBF)	67.92		58.28	**74.06**
	SHLA	67.67		58.96	70.23
EMOVO	Direct C	58.02	51.25		55.31
	Direct SC	59.10	55.42		70.94
	KCCA (Lin.)	62.92	**58.54**		65.94
	KCCA (RBF)	60.23	56.04		**71.87**
	KPCA (Lin.)	65.38	56.04		62.81
	KPCA (RBF)	66.70	56.04		67.50
	SHLA	**67.32**	58.16		64.32
Polish	Direct C	65.09	55.42	57.29	
	Direct SC	65.92	56.87	54.32	
	KCCA (Lin.)	70.93	58.75	56.10	
	KCCA (RBF)	68.45	**61.87**	54.12	
	KPCA (Lin.)	69.33	58.75	57.79	
	KPCA (RBF)	67.91	57.08	57.64	
	SHLA	**71.10**	60.57	**58.27**	

Moreover, using KCCA prevents the necessity of having the same number and type of features. On the other hand, autoencoders with large number of layers and nodes with non-linear activation function can represent better non-linearity in the data distribution. However, to achieve this non-linear mapping there is a need of much more data samples.

3.4 Discussion and future work

We improved the cross-lingual emotion recognition using Kernel Canonical Correlation Analysis (KCCA) on principal component subspaces to increase the similarity between the two source and target corpora. We compared this approach with the Kernel PCA as well as a state-of-the-art autoencoder approach named as Shared Hidden Layer Autoencoder (SHLA). On average, the proposed approach performs better than the others.

The future study is toward the use of non-linear mapping instead of linear PCA to generate subspaces. Non-linear mapping can be achieved through Deep Canonical Correlation Analysis where Deep Neural Networks are trained to reduce the dissimilarity between two distributions [56], [57]. Additionally, discriminant mappings (such as Linear Discriminant Analysis [58]) could be applied on the training corpora to increase the level of discrimination on the corresponding subspace.

4 Conclusion

In the MixedEmotions project, emotion recognition system should be universal and ready to cope with the multi-lingual audio data. In this respect, due to unavailability of annotated data, it is infeasible to create models for each language. Therefore, in this deliverable we investigated the cross-lingual emotion analysis, assuming we have annotation for one language and we want to recognize emotions in audios of another language. First, we found that having source and target from the "same language," we achieve the highest accuracy and then choosing the source language in the "same language family" yields higher accuracy than "between language families." We further improved the cross-lingual emotion recognition using transfer learning. The results demonstrate the possible methodologies to cope with the multilingual aspect of the MixedEmotions project where none or few annotations are provided for a data source.

5 References

[1] J. J. Ohala and others, "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," *J. Phon.*, vol. 37, no. 4, pp. 417–435, 2009.

[4] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *Int. J. Speech Technol.*, 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, 2013, pp. 1–5.
[7] P. Fung and T. Schultz, "Multilingual spoken language processing," *Signal Process. Mag. 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," *J. Mem. Lang.*, 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," *J. Cross. Cult. Psychol.*, vol. 32, no. 1, pp. 76–92, 2001.
[10] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in Cantonese and English," *J. Acoust. Soc. Am.*, 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," *role prosody Affect. 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," *Pragmat. Cogn.*, vol. 19, no. 2, pp. 203–223, 2011.
[13] 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.
[14] H. R. Markus and S. Kitayama, "The cultural construction of self and emotion: Implications for social behavior," *Emot. Soc. Psychol. Essent. Read.*, pp. 119–137, 2001.
[15] S. P. Whiteside, "Acoustic characteristics of vocal emotions simulated by actors," *Percept. Mot. Skills*, vol. 89, no. 3f, pp. 1195–1208, 1999.
[16] I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database des," *Intern. AAU report, Cent. Pers. Kommun. Denmark*, p. 22, 1996.
[17] 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*, 2003, pp. 85–89.
[18] 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.
[19] 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*, 2015.
[20] 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.
[21] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions," in *LREC*, 2006, vol. 6, pp. 15–19.
[22] S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim, M. Tu, M. Intelligence, S. Processing, B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge□: Social Signals , Conflict , Emotion , Autism □," *Interspeech2013*, pp. 148–152, Aug. 2013.
[23] Å. Abelin and J. Allwood, "Cross linguistic interpretation of emotional prosody," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.
[24] V. Makarova and V. A. Petrushin, "Phonetics of emotion in Russian speech," in *XVth international conference of phonetic sciences*, 2003.
[25] 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.
[26] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. of ASRU*, 2009.
[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.
[28] 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*, 2006, p. 8.
[29] 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*, 1998, vol. 98, pp. 923–926.
[30] S. M. Feraru, H. N. Teodorescu, and M. D. Zbancioc, "SRoL-Web-based resources for languages and

language technology e-Learning," *Int. J. Comput. Commun. Control*, vol. 5, no. 3, pp. 301–313, 2010.
[31] H.-N. Teodorescu and S. M. Feraru, "A study on Speech with Manifest Emotions," in *Text, Speech and Dialogue*, 2007, pp. 254–261.
[32] H. K. Ekenel, M. H. Meral, and S. A. Ozsoy, "Analysis of Emotion in Turkish," in *XVII.| National Conference on Turkish Linguistics*, 2003.
[33] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.
[34] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2011 - The First International Audio/Visual Emotion Challenge," in *Proc. of the 1st International Audio/Visual Emotion Challenge and Workshop, AVEC 2011, held in conjunction with the International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2011, ACII 2011*, 2012, vol. II, pp. 415–424.
[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.
[36] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9, pp. 1062–1087, 2011.
[37] 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*, 2013, pp. 511–516.
[38] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, 2014, no. Avec 2012, pp. 3592–3598.
[39] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.
[40] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," *Inf. Fusion*, vol. 24, pp. 84–92, 2015.
[41] a. Gretton, a. J. Smola, J. Huang, M. Schmittfull, K. M. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift Mach. Learn.*, vol. 3, no. 4, p. 5, 2009.
[42] A. Hassan, "On automatic emotion classification using acoustic features," University of Southampton, 2012.
[43] J. Deng, R. Xia, Z. Zhang, and Y. Liu, "Introducing Shared-Hidden-Layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014, vol. 338164, no. 338164, pp. 4851–4855.
[44] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.
[45] D. R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fMRI data using kernel canonical correlation," *Neuroimage*, vol. 37, no. 4, pp. 1250–1259, 2007.
[46] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 82–86.
[47] H. Kaya, F. Eyben, A. A. Salah, and B. B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 3729–3733, 2014.
[48] Y. Li and J. Shawe-Taylor, "Using KCCA for Japanese--English cross-language information retrieval and document classification," *J. Intell. Inf. Syst.*, vol. 27, no. 2, pp. 117–133, 2006.
[49] B. Fortuna and J. Shawe-Taylor, "The use of machine translation tools for cross-lingual text mining," in *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.
[50] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[51] S. Haq, P. J. B. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP), Tangalooma, Australia*, 2008.
[52] G. Costantini, I. Iaderola, andrea Paoloni, and M. Todisco, "EMOVO Corpus: an Italian Emotional Speech Database," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014.
[53] P. Staroniewicz and W. Majewski, "Polish Emotional Speech Database – Recording and Preliminary Validation," in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, vol. 5641,

Springer Berlin Heidelberg, 2009, pp. 42–49.
[54] M. I. Retrieval and S. Processing, "openSMILE□:) The Munich Open-Source Large-scale Multimedia Feature Extractor," vol. 6, no. 4, pp. 4–13, 2014.
[55] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, 2010.
[56] W. Wang, K. Livescu, and J. Bilmes, "On Deep Multi-View Representation Learning," *Icml*, vol. 37, 2015.
[57] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, pp. 1247–1255, 2013.
[58] M. Kan, S. Shan, H. Zhang, and X. Chen, "Multi-View Discriminant Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, 2016.