On Acoustic Emotion Recognition: Compensating for Covariate Shift – Supplementary Results

Ali Hassan, Bob Damper, Mahesan Niranjin

Electronics and Computer Science
University of Southampton

April 23, 2012

Abstract

This report contains the supplementary material for the paper titled 'On Acoustic Emotion Recognition: Compensating for Covariate Shift' which has been submitted to IEEE Transactions on Audio Speech and Language Processing. This report contains the SD-CV, SI-CV and inter-database results on three commonly used acted emotional speech databases.

1 Introduction to Acted Databases

Before we start discussing the results on the three freely available acted databases, first we give some details of these datasets.

1.1 Danish Emotional Speech Database

The Danish emotional speech (DES) database is described in [1]. It is only available for non-commercial research use. DES was recorded in Aarhus Theatre for Center for Person Komunikation (CPK), Aalborg University, Denmark in 1995. Four professional speakers, 2 males and 2 females, were asked to speak predefined sentences and words in Danish for 5 emotions: *neutral*, *angry*, *happy*, *sad* and *surprised*. Each speaker was asked to say 2 words, 9 short sentences and 2 passages ('paragraphs') in all 5 emotions. The average length of spoken words is 1 s; the sentences consist of on average 4.5 words lasting for 1.5 s. The paragraphs consist of 2 and 4 sentences each lasting for 10 s and 26 s, respectively. A total of 260 sentences is available in the database, with 52 sentences per emotion class making up 28 minutes of speech material. All recorded samples were included in the database. The quality of the acted emotions was verified by 20 human listeners, who were allowed to listen to them as many times as they wished before classifying them into one of the five emotion classes. This revealed that the *neutral* emotion is very strongly confused with *sad*; *angry* with *neutral* and *surprised*; *happy* with *neutral* and *surprised*; and *surprised* with *happy* and *neutral*. Reported human accuracy on this database is 67.3%.

Other researchers have treated the two passages differently. Sometimes they are left out of the training and testing sets, whereas in other cases they are divided into sentences (by detecting inter-sentence pauses) leading to a database consisting of over 400 sentences. In our work, to keep things simple and make future comparisons easier, we have omitted the passages.

Table 1: Emotion classes and number of sentences per class for acted DES, Berlin and Serbian databases. The horizontal line in table separates the emotions that are common to all three acted databases from those which are not.

DES	**Sentences**	**Berlin**	**Sentences**	**Serbian**	**Sentences**
Neutral	52	Neutral	79	Neutral	558
Angry	52	Anger	127	Anger	558
Happy	52	Happiness	71	Happiness	558
Sad	52	Sadness	62	Sadness	558
Surprised	52	Fear	69	Fear	558
		Boredom	81		
		Disgust	46		
Speakers	2M,2F		5M, 5F		3M, 3F

1.2 Berlin Database

The Berlin database [2], also known as Emo-DB, contains utterances spoken in German. It is available at `http://pascal.kgw.tu-berlin.de/emodb/index-1024.html` (last visited 17 Apr 2012). The database was recorded in 1997 and 1999 in an anechoic chamber at the Technical University, Berlin. Ten professional native German actors, 5 males and 5 females, were asked to speak 10 sentences in 7 different emotions: *neutral*, *anger*, *happiness*, *sadness*, *fear*, *boredom* and *disgust*. Note that four of these classes are common with DES. These sentences were then evaluated by 20–30 listeners to verify the emotional state and only those were retained that had a recognition rate of 80% or above and were judged as natural by more than 60% of the listeners, yielding "about 500 utterances" in total making up 22 minutes of speech material. Each sentence consists of on average 10 words with average duration of approximately 5 s. Reported human accuracy on this database is 86.1%.

1.3 Serbian Database

The Serbian database of acted emotional speech [3] was recorded in 2003 in an anechoic studio at the Faculty of Dramatic Arts, Belgrade University, Serbia, using 6 actors: 3 males and 3 females. It has been less well used than DES and Berlin. It consists of 32 isolated words, 30 short semantically-neutral sentences, 30 long semantically-neutral sentences and one passage consisting of 79 words, i.e., $32 + 30 + 30 + 1 = 93$ utterances. The following 5 emotions are represented: *neutral*, *anger*, *happiness*, *sadness* and *fear*. Hence, there are $93 \times 6 = 558$ sentences per emotion. Each of the 93 utterances is contained in a separate `.wav` file; so there are $93 \times 6 \times 5 = 2790$ files in total. Each speaker was recorded in separate sessions so that they do not influence each other's speaking style. Each recorded sentence was evaluated by 39 listeners; reported human accuracy on this database is 94.7%. In general, these human listening tests show that *anger* and *happy* emotions are often confused with each other, whereas *neutral* is most usually confused with *sad*.

2 Results of K-S Tests

To verify the existence of covariate shift, we have proposed to apply Kolmogrov–Smirnov (K-S) test in different scenarios. This test is applied on the corresponding features from the training and testing data. Table 2 shows the percentage average out of 6552 features failing the test.

Table 2: Average percentage out of 6552 features failing the K-S test.

	% of features		
Method	**DES**	**Berlin**	**Serbian**
SD-CV	4.8	6.7	4.5
SI-CV	35.1	37.6	77.3

Table 3: Mapping of emotional classes for the three acted speech databases on Arousal and Valence dimension.

	Arousal			
Database	**Low**	#	**High**	#
DES	Neutral, Sad	156	Angry, Happy, Surprised	104
Berlin	Boredom, Disgust, Neutral, Sad	267	Angry, Fear, Happy	268
Serbian	Neutral, Sad	1674	Angry, Fear, Happy	1116
	Valence			
	Negative	#	**Positive**	#
DES	Angry, Sad	156	Happy, Neutral, Surprised	104
Berlin	Angry, Boredom, Disgust, Fear, Sad	150	Happy, Neutral	385
Serbian	Angry, Fear, Sad	1116	Happy, Neutral	1674

3 Mapping of Emotion Classes

It has been discussed before that each emotional speech database has a different number of classes per database. Hence, we can not directly apply inter-database classification. One solution is to apply inter-database emotion classification on only the common classes between all of the databases. The three acted emotional speech database (DES, Berlin and Serbian) have four classes common among each other. These classes are *neutral*, *angry*, *happy* and *sad* (refer to Table 1).

Another solution is to map all classes on a lower dimensional space. For doing this mapping, *valence* and *arousal* dimensions are our best options. We choose these dimensions as these two are usually considered as the two basic dimensions to represent any emotions by the dimensional theory for emotions. Testing on these dimensions individually will give us further insight into their representation in the data. We expect that classification accuracy for *valence* will be significantly lower than *arousal* dimension.

As the labels for these dimensions are not available, we use the circumplex model of affect for speech to map the corresponding emotion classes to these two dimensions. We map all emotions in the corresponding databases to low or high *arousal* and negative or positive levels of *valence*. This mapping of emotions and the number of samples per class for these acted emotional speech databases are shown in Table 3.

4 SD-CV and SI-CV Results of the Mapped and Common Classes

To establish the baseline results on these databases, we apply SD-CV and SI-CV on these databases individually. Same setup is used as is mentioned in the paper. The results of applying SD-CV and SI-CV classification for *arousal*, *valence* and 4 common classes among all of the acted databases are given in Table 4(a) and Table 4(b) respectively. On average, we get 97.8% and 84.0% SD-CV UA accuracy for *arousal* and *valence* dimensions respectively. From these results it is clear that the *arousal* dimension is much easier to recognise as compared to the valence dimension. Worst results for *valence* recognition are obtained for the DES database (74.5% UA) while for the other two, they are above 90% UA which is very good. This means that for this database, it is not only difficult to separate *angry* from *happy* which have positive *valence*, but these two are also not very easily separable from *neutral* and *sad* emotions in the *valence* dimension. For four common classes among the three database, best accuracy is obtained for the Serbian database (91.6% UA) while worst results are obtained for the DES database (76.0% UA).

Interestingly, the classification results for SI-CV are very close to SD-CV especially by using CMN+MLLR and IW-algorithms. In some of the cases (DES and Berlin) by using these algorithms we get very large improvements in comparison to using the standard SVM classifier. This actually fits with the theoretical basis of these methods as there is a larger room for improvements for SI-CV than for SD-CV, which is seen from the results.

An important observation is that the average results for *arousal* and *valence* recognition by CMN+MLLR and IW-algorithms for all of the database are better than the results of standard linear SVM. This means that by using methods that explicitly compensate for the speaker and environmental differences improve the results significantly.

The CMN+MLLR algorithm does improve the classification results in comparison to the standard SVM. However, when compared againt the three IW-algorithms, it only performs better in 1 out 18 SI-CV and SD-CV experiments. Generally, we get better results by applying IW-algorithms which compensate for the covariate shift in the data. Out of the three IW-algorithms, uLSIF performs best in 7 out of 18 experiments. This shows that just like CMN+MLLR, IW-algorithms can also be successfully used to compensate for the mismatch between the training and testing data caused by different speakers.

5 Inter-Database Classification of the Mapped and Common Classes

The results of inter-database emotion classification are given in Table 5. They are obtained by applying leave-one-database-out cross validation. The database marked at the top of each column was used for testing while the remaining two were used for training the classifiers. It can be observed that inter-database accuracy for *arousal*, *valence* and four common classes is lower than intra–database classification accuracy. This is very much expected as the recording environments and speakers for the training and testing data are separate and different from each other. This kind of situation is the one which will be faced by any practical SER system. In such a situation, one has to apply some methods to compensate for the mismatch. From the results shown in Table 5, one can see that CMN+MLLR does significantly increase the classification accuracy as compared to standard SVM. However, increase in the classification accuracy is less as compared to the IW-algorithms. Out of the three IW-algorithms, uLSIF based classification performs best in 7 out of 9 experiments. Hence, we declare uLSIF as the best out of the three tested IW-algorithms.

Table 4: SD-CV and SI-CV intra-database percentage UA accuracy on three acted databases for *arousal*, *valence* and 4-common classes using traditional CMN+MLLR method and the three IW-algorithms from transfer learning. The numbers in the brackets are the standard deviations.

(a) SD-CV Intra-database classification results.

	DES			**Berlin**			**Serbian**		
Method	**Arousal**	**Valence**	**4-Class**	**Arousal**	**Valence**	**4-Class**	**Arousal**	**Valence**	**4-Class**
SVM	95.5 (3.2)	71.8 (7.8)	74.6 (8.7)	95.9 (2.7)	93.6 (2.8)	84.8 (4.1)	99.5 (0.3)	91.2 (0.9)	91.3 (2.0)
CMN+MLLR	97.0 (3.0)	70.5 (6.4)	74.7 (8.5)	96.6 (4.6)	92.9 (2.7)	84.9 (4.1)	**99.5** (0.1)	91.0 (1.1)	91.1 (2.6)
KMM	**99.2** (1.9)	75.2 (6.6)	76.9 (10.0)	97.2 (3.3)	93.3 (3.1)	85.2 (5.0)	98.1 (0.7)	**94.4** (0.8)	91.8 (2.22)
KLIEP	97.4 (4.7)	74.1 (8.4)	76.4 (9.7)	97.2 (2.7)	**93.9** (2.0)	**87.6** (2.6)	99.5 (0.4)	91.4 (0.5)	91.6 (3.6)
uLSIF	97.2 (3.0)	**75.8** (8.5)	**77.2** (7.1)	**97.4** (3.4)	92.1 (1.5)	86.0 (4.1)	99.5 (0.4)	91.1 (1.1)	**92.1** (1.9)
Mean	**97.3**	**74.5**	**76.0**	**96.8**	**93.2**	**85.7**	**99.2**	**91.8**	**91.6**

(b) SI-CV Intra-database classification results.

	DES			**Berlin**			**Serbian**		
Method	**Arousal**	**Valence**	**4-Class**	**Arousal**	**Valence**	**4-Class**	**Arousal**	**Valence**	**4-Class**
SVM	88.6 (3.9)	76.5 (7.6)	76.0 (8.6)	93.3 (6.1)	92.1 (1.5)	84.8 (2.7)	96.4 (3.9)	88.5 (2.6)	81.2 (7.5)
CMN+MLLR	89.0 (4.5)	77.1 (9.3)	76.0 (7.5)	94.8 (3.0)	92.3 (3.1)	87.5 (4.1)	96.2 (2.4)	90.5 (3.2)	83.5 (6.9)
KMM	**91.5** (8.3)	**83.1** (9.0)	77.9 (11.5)	98.3 (1.5)	93.3 (3.0)	91.6 (1.7)	**97.6** (2.8)	91.6 (3.8)	84.1 (6.6)
KLIEP	91.3 (6.5)	82.6 (8.5)	76.4 (3.7)	97.9 (1.5)	**93.9** (2.4)	**92.3** (1.7)	96.9 (3.5)	90.6 (2.3)	**84.7** (6.1)
uLSIF	90.0 (6.7)	81.1 (6.9)	**78.8** (7.1)	**98.3** (1.2)	93.6 (2.8)	89.1 (1.8)	97.0 (1.9)	**91.7** (5.3)	84.1 (6.5)
Mean	**90.1**	**80.1**	**77.0**	**96.5**	**93.0**	**89.1**	**96.8**	**90.6**	**83.5**

Table 5: Inter–database percentage UA accuracy on three acted databases for *arousal*, *valence* and 4-common classes using traditional CMN+MLLR method and the three IW-algorithms from transfer learning.

Testing on →	**DES**			**Berlin**			**Serbian**		
Method	**Arousal**	**Valence**	**4-Class**	**Arousal**	**Valence**	**4-Class**	**Arousal**	**Valence**	**4-Class**
SVM	71.4	50.8	40.5	72.9	49.2	39.5	82.7	64.2	63.3
CMN+MLLR	74.1	50.4	41.3	74.6	50.0	40.0	83.9	67.4	65.0
KMM	75.0	51.5	42.8	75.0	50.1	43.6	84.9	66.7	**65.5**
KLIEP	75.4	51.5	43.3	73.3	58.2	41.3	85.8	**69.3**	65.2
uLSIF	**82.4**	**51.7**	**44.7**	**75.8**	**58.4**	**46.1**	**87.8**	69.1	64.5
Mean	**75.7**	**51.2**	**42.5**	**74.3**	**53.2**	**42.1**	**85.0**	**67.3**	**64.7**

These results are very interesting as all of the three databases tested are in different languages. Although German and Danish belong to the same family of Germanic languages and thus share some similarities. The Serbian is a Slavonic language which does not belong to the same family. On average, we get 78.3% and 57.3% UA accuracies for *arousal* and *valence* recognition by testing on the database which has different speakers, recording environments and different language than those used for training the classifiers. These are very good results considering such large differences between the training and testing datasets. Especially, the UA for inter-database *arousal* recognition is very high and UA accuracy for inter-database *valence* recognition is also above chance level. Best inter-database classification results are obtained for testing on the Serbian database. As mentioned earlier, this database does not belong to the family of Germanic languages so the expected results should have been opposite. However, average accuracy on this database is generally very high which is the reason for these results. Secondly, all of these databases contain European languages. So there are some cultural aspects common between them. These arguments can explain these results.

These experiments show that there are some aspects of emotions which are *universal across several languages*. Even if the classifier does not have any information about the test language, it can still get quite reasonable results, better than random guessing. These results also validate our assumption that by using different databases for training and testing, which have different speakers, acoustic environments and languages as well, introduces a shift in the data which can be compensated by traditional methods used in ASR systems as well as IW-algorithms. Generally, IW-algorithms perform better than CMN+MLLR, and out of the three algorithms tested, uLSIF performs the best.

References

[1] I. Engberg and A. Hansen, *Documentation of the Danish Emotional Speech Database DES*, Center for PersonKommunikation, Institute of Electronic Systems, Alborg University, Aalborg, Denmark, 1996.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of 9th European Conference on Speech Communication and Technology, Interspeech'05*, Lisbon, Portugal, 2005, pp. 1517–1520.

[3] S. T. Jovicic, Z. Kacic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: Design, processing and evaluation," in *Proceedings of 9th Conference on Speech and Computer, SPECOM'04*, St. Petersburg, Russia, 2004, pp. 77–81.