Five Attempts at Cross-Dataset Speech Emotion Recognition

Andrejaana Andova
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
andrejaana.andova@ijs.si

Mitja Luštrek
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
mitja.lustrek@ijs.si

ABSTRACT

In this paper, we tried to recognize emotions from speech when we have a very small dataset available. Because recording and annotating new data is a costly task, our goal is to use publicly available datasets in addition to our own to improve the recognition accuracy. To do this, we implemented five different methods able to extract knowledge from the publicly available dataset and use it in our target dataset. Two of these methods are based on transfer learning, one is a multi-task learning method, and the last two use advanced feature normalization methods to bring the feature domains from the two datasets closer together. We show that in certain combinations of train-test set, some of our methods outperform the baseline classifier by a maximum of 9 percentage points. In some cases, however, the baseline method proved to provide best results.

Keywords

speech emotion recognition, transfer learning, multi-task learning, cross-dataset

1. INTRODUCTION

In the last two decades a lot of research has been put into the automatic recognition of emotional states. The main reason for this is the rapid development of affective user interfaces. If we are able to recognize the user's emotions, we can develop dynamic applications that can adapt to the user's feelings at any given time. Here, our task could be to measure the stress of the calling users and include them in a priority list accordingly.

In our research, we focused on measuring the emotional state of a person based on their speech. More precisely, we did not analyse what the person is saying, but how they express themself. Since this type of emotion recognition does not analyse the content of the person's speech and does not require any of their other personal data, it can be used if we have some limitations regarding the use of personal data.

The main obstacle to determining a person's emotional state for machine learning is the lack of data. Bringing people into a certain emotional state is a challenging and, when it comes to negative emotions, unethical task. Furthermore, recording people expressing their genuine feelings, without their knowledge, violates the right to privacy, and these datasets, if ever acquired, are not publicly available. Therefore, most of the datasets we get are from actors who do their best to perform certain emotions.

Many papers present high accuracy scores when training and testing a model with the same dataset [9, 6, 10] but, what happens if we do not have data for our specific problem? What happens if we use recording devices of different quality, or if our target subjects are older people and not young or middle-aged actors, or if our setting differs in some other way? In our research, we assumed that we have a small amount of representative data of our problem, and we wanted to additionally exploit publicly available datasets. The most similar related work is a contribution by Latif et al. from INTERSPEECH 2018 [8]. Although it claims to achieve state-of-the-art results, when trained on a source dataset and tested on a target its accuracy does not even exceed the majority classifier. When a small amount of data from the target dataset was used, a modest improvement was observed, but the paper only distinguishes between two emotions. It thus seems that the issue we are tackling in this paper is poorly explored in the literature.

The most common way to use information from a source dataset to improve the classifier of a target dataset is to use transfer learning. We have chosen two different methods for transfer learning. The first method was recreated from the paper mentioned previously, while the other method uses Fully Connected Neural Networks to transfer some of network parameters between datasets. Additionally, we tried multi-task learning as well as two different types of feature normalization, which we applied on the data in order to bring the feature domains between the two datasets closer together.

In Section 2 we present the datasets we use. In Section 3 we present the five methods we used for cross-dataset emotion recognition. In Section 4 we present our evaluation methods and the achieved results. Finally in Section 5, we conclude and present our future work.

2. DATASETS

To detect emotions from speech, we used four publicly available datasets: EmoDB [1], EMOVO [3], IEMOCAP [2] and SAVEE [7]. The datasets were recorded in three different languages: IEMOCAP and SAVEE datasets were recorded in English, while EMOVO and EmoDB were recorded in Italian and German. A common problem when combining multiple datasets is that most of the datasets use different sets of emotions. To deal with this problem, we used only instances presenting four basic types of emotions, which are present in all the datasets: neutral, anger, joy and sadness. The number of instances for each emotion per dataset is presented in Table 1.

Table 1: Number of instances per emotion

Dataset	Neutral	Anger	Joy	Sadness
EmoDB	79	127	71	62
EMOVO	84	84	84	84
IEMOCAP	392	500	94	467
SAVEE	120	60	60	60

3. METHODS

The best established way to build a model able to recognize emotions from speech is by extracting global features from the speech and then building a classifier on top of these features. To extract features, we used a publicly available toolkit - OpenSmile [4], which offers a wide range of possible sets of features. We decided to use the 'emobase2010' feature set. This set is composed of overall 1582 features. As the machine learning algorithm we selected Random Forest with 1000 trees and maximal depth of 10. This combination outperformed several alternatives, including Deep Learning on raw audio. An additional advantage of the OpenSmile features and Random Forest over Deep Learning is that features can easily be extracted on the phone, so that raw audio is never sent outside the user's device. We developed or implemented five methods for transfer learning, which we describe in sections 3.1-3.5.

3.1 Deep Belief Network

In speech emotion recognition, there has not been much related work whose main focus is to transfer knowledge from the source dataset to the target dataset. The most dedicated attempt is the already mentioned one by Latif et al. [8], in which they used Deep Belief Network (DBN). To evaluate their method, they used another work that used autoencoders to transfer knowledge from the source dataset to the target dataset. They achieved better results than the autoencoders approach, and therefore we decided to recreate their method. In our DBN implementation, we used the same network parameters as described in their paper.

3.2 Fully Connected Deep Neural Network

Since the rise of transfer learning, the most commonly used method of transferring knowledge from one problem in another is by transferring network parameters. Therefore, in the second method we trained a Fully Connected Neural Network (FCNN) on the source dataset and transfered some of the network parameters to the target dataset. The FCNN architecture is composed of one input layer, one output layer, and three hidden layers. The input layer takes the same amount of input units as the number of features extracted from one utterance. The first hidden layer is composed of 1000 units, the second hidden layer is composed of 500 units and the third hidden layer is composed of 300 units. The output layer consists of only four units, one for each emotion. The activation function of all layers is 'tanh'. The only exception is the output layer, which uses 'softmax' activation function.

First, the FCNN was trained on the whole source dataset. After the training on the source dataset was finished, and all network parameters have been determined, we froze all parameters of the network, except those belonging to the output layer. We then fine-tuned the parameters of the final layer using a part of the instances from the target dataset.

3.3 Multi-task Learning

In the multi-task learning method, we used the same Random Forest classifier as the one described in the baseline method. However, instead of having the same target class for matching emotions from the source and the target dataset, we used two different target classes: one for the emotion from the source dataset, and another one for the same emotion from the target dataset. For example, angry utterances from the source and the target dataset would get the same target label 'Anger' in the baseline Random Forest Classifier. However, in the multi-task learning approach, angry utterances from the source dataset would get the target label 'Anger1', while angry utterances from the target dataset would get the target label 'Anger2'. The idea is that the classifier classifies specifically into classes of the target dataset, while the structure of the classifier still benefits from the source dataset (upper leaves of the tree in Random Forest). Because when training we used the whole source dataset and only a small portion from the target dataset, we ended up with unequal distribution of emotions. To deal with this problem, we oversampled examples from the target dataset until we got equal distributions in both datasets.

3.4 Normalization based on neutral speech

As shown in Table 1, the distribution of emotions is not equal across dataset and thus, a simple feature normalization and standardization method might not work across different datasets.

To implement a more advanced feature normalization method, we applied a normalization technique on the source and on the target data independently. In this normalization technique, we used neutral speech to bring the datasets to the same reference point. Ideally the neutral speech using our normalization technique should be near the coordinate space origin. To normalize and standardize our data, we applied the following formula to the feature values:

$$x_{i\_new} = \frac{x_i - \mu}{\sigma}$$

where μ is the average value from neutral speech in the training data, σ is the standard deviation in all training data and x_i is the i-th instance in the data.

To evaluate the performance of the model whose features were normalized based on the neutral emotion, we used the baseline Random Forest Classifier.

3.5 Normalization based on feature distribution

When analysing feature distributions between two datasets, we noticed that most of the features do not have the same distributions per emotion. For example, on the left side in Figure 1 we present the distribution of feature 'pcm_loudness_sma_amean' on neutral utterances in IEMOCAP, while on the right side, the distribution of the same feature is presented for neutral utterances in SAVEE. Since this could be confusing for our model, we tried to bring the two feature distributions as close as possible for each emotion.

To do this, we used the feature distribution from the training data in the target dataset as the baseline, and tried to bring the feature distribution from the whole source dataset as close as possible to the baseline distribution. Thus, for each emotion, we divided the feature distribution of the training data in the target dataset into 5 equal bins and saw how

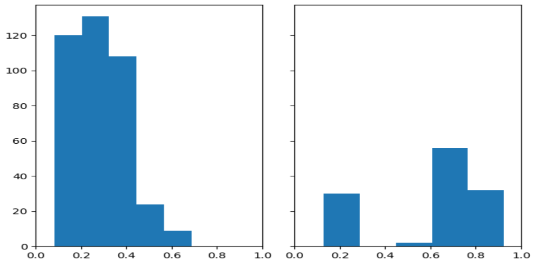

Figure 1: Distribution of values in neutral speech from feature pcm_loudness_sma_amean on the left side in IEMOCAP, on the right side for SAVEE

much of the data belongs to each bin. Thus, all instances in the target dataset whose $feature_i$ belong to the first bin of the $feature_i$ distribution, sould be given the value 0, all instances that belong to the second bin of the $feature_i$ distribution would be given the value 1, etc.

Let us assume that the percentage of the training data from the target dataset that was given the value 0 for $feature_i$ is x_0, the percentage of the training data that was given the value 1 is x_1, etc. To bring the distributions for $feature_i$ from the source and the target dataset closer together, we would assign the lowest x_0 percent of data from the source dataset a value of 0, etc. Thus, we got a similar distribution of the features in the source and the target dataset for each emotion separately.

To evaluate the performance of the model whose features were normalized based on their distribution, we used the baseline Random Forest Classifier.

4. EVALUATION AND RESULTS

Recording and annotating data is an expensive process. Because of this, for training our models we used at most one subject, while the rest of the subjects would be used for testing the performance of the model. To evaluate the performance of the model on the target dataset without a transfer-learning method, we used three different scenarios:

- In the first scenario, we uses the whole source dataset for training, and the whole target dataset for testing. This simulates the case when we do not have any of our target data, thus only training our model on a publicly available dataset.
- In the second scenario, we used one subject from our target dataset as the training set and the rest of the subjects as the testing set. To get a more objective evaluation of the performance of the model, each of the subjects was used as training data once, and the final result was calculated by averaging the accuracy of each train-test split. This scenario simulates the case where we only use our small dataset for training.
- In the third scenario we used the whole source dataset and one subject from the target dataset for training. The rest of the subjects from the target dataset were used for testing. Similarly as in the second scenario, each subject from the taget dataset was used as the training data once, making the final result an average of each train-test split. This scenario simulates the case where we combine our small dataset with some publicly available dataset.

Since only two of the four datasets we use are recorded in the same language, we decided to evaluate the models only on these two datasets. This way, we can compare same-language and different-language cross-dataset emotion recognition on the same target datasets. To present the complexity of the task, in Table 2 we used the Majority Classifier and Random Forest Classifier for each of the three possible scenarios. The results achieved using the first scenario are the poorest, not even achieving the majority classifier. The results achieved using the second and the third scenario outperform the majority classifier, but still the classifier from the third scenario gives overall poorer results compared to the classifier from the second scenario. This could mean that we do not gain any useful information from the source dataset.

We applied the transfer learning methods from Section 3 according to the third scenario. The results are presented in Table 3. To evaluate the success of the information transfer, we compared these results to the baseline Random Forest Classifier calculated using the second scenario. The results in Table 3 show us that most of the improvements are achieved by normalizing the feature spaces based on the feature distribution. However, the presented results are not optimistic, since in some cases the best results were achieved using the baseline classifier. So far, the best improvement we achieved was 9 percentage points, which we gained when training on EmoDB and testing on SAVEE while normalizing the features based on their distributions. This method outperformed both the DBN presented as most suitable for this type of problems in related work, as well as the commonly used FCNN transfer learning.

An interesting observation is that when our methods use EmoDB and EMOVO as train data and SAVEE as test data, they perform better compared to when the same-language IEMOCAP is used as train data. This happened with most of our methods, and could indicate that the way the recording took place (5 min conversations vs. short utterances), might be more important when choosing which source dataset to use, than the language.

5. CONCLUSIONS

In this paper, we tried to use the knowledge obtained from a source dataset in order to improve the classification accuracy of a target dataset. We found that although in different languages, EmoDB and EMOVO contain more useful information for detecting emotions from speech in SAVEE, compared to the same-language database IEMOCAP.

The baseline classifier could be outperformed by using some of the methods described here, with a maximum improvement of 9 percentage points. The best performance was achieved by normalizing the features, based on their distributions. The worst performance was achieved by a method from related work, which did not even outperform the majority classifier.

Although we implemented five different methods for cross-dataset speech emotion recognition, there are other possibilities. A potentially more effective, but substantially more

Table 2: Results obtained from the majority classifier and baseline Random Forest Classifier for each scenario without transfer learning

Train dataset	Test dataset	Majority	Scenario1	Scenario2	Scenario3
EmoDB	SAVEE	40%	29%	49%	57%
EMOVO	SAVEE	40%	41%	49%	51%
IEMOCAP	SAVEE	40%	27%	49%	41%
EmoDB	IEMOCAP	34%	34%	67%	62%
EMOVO	IEMOCAP	34%	52%	67%	67%
SAVEE	IEMOCAP	34%	33%	67%	65%

Table 3: Results obtained from the majority classifier and baseline Random Forest Classifier compared to the five transfer learning methods

Train dataset	Test dataset	Majority	Baseline RF	DBN	FCNN	Multi-task	Norm.1	Norm.2
EmoDB	SAVEE	40%	49%	20%	50%	56%	57%	58%
EMOVO	SAVEE	40%	49%	30%	58%	50%	51%	58%
IEMOCAP	SAVEE	40%	49%	20%	46%	42%	41%	47%
EmoDB	IEMOCAP	34%	67%	33%	60%	62%	62%	66%
EMOVO	IEMOCAP	34%	67%	30%	61%	67%	67%	67%
SAVEE	IEMOCAP	34%	67%	27%	56%	66%	65%	69%

complex approach may be by using Generative Adversarial Networks [5] to translate the features from one dataset to another. This is the main approach we consider for future work, assuming it can work on modestly sized dataset such as are available.

A cursory look at the literature suggests that emotion recognition from speech is not a very difficult problem, since many papers report good results and several datasets are publicly available. However, our study shows that practical applicability of these datasets is limited considering how poorly cross-dataset learning works. It is also possible that the typical methods for emotion recognition from speech would prove unsuitable for the wider range of emotion expressed in real life. Therefore, it is important to study emotion recognition without limiting to one homogeneous dataset.

6. ACKNOWLEDGEMENTS

We would like to thank our colleagues Benjamin Fele and Aljoša Vodopija for their help in this research.

This Research has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 769765

7. REFERENCES

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[3] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco. Emovo corpus: an italian emotional speech database. In *International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3501–3504. European Language Resources Association (ELRA), 2014.

[4] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*, 2014.

[7] S. Haq, P. J. Jackson, and J. Edge. Audio-visual feature selection and reduction for emotion classification. In *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*, 2008.

[8] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps. Transfer learning for improving speech emotion classification accuracy. *arXiv preprint arXiv:1801.06353*, 2018.

[9] T. L. Nwe, S. W. Foo, and L. C. De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[10] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.