Improving Speech Emotion Recognition by Identifying the Speaker with Multi-task Learning

Mei Lan Schrama
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
m.l.schrama@student.utwente.nl

ABSTRACT

Automatic emotion recognition is an important topic in artificial intelligence as it can improve our experience when interacting with machines. This paper proposes a multi-task learning (MTL) algorithm for speech emotion recognition as the main task and speaker identification as the auxiliary task. After experimenting both separate training and MTL training on the same dataset, the proposed method results in a relative improvement of 15.6% on the accuracy of emotion recognition. Additionally, it is observed that the MTL model does not make a contribution on the correctness of identifying speakers.

Keywords

Speech emotion recognition, Speaker identification, Multi-task learning.

1. INTRODUCTION

Speech emotion recognition (SER) has drawn great attention from scientists worldwide. SER aims to recognize the emotion of speakers by analysing the vocal features of the speech such as pitch, energy and other segment level features. During the development of SER techniques, scientists have applied a diversity of methods to increase the accuracy of the recognition. A direction to improve the performance of machine learning is to achieve a better generalization, which refers to the ability of the machine to predict correct outcome for unseen data. This is a common problem in machine learning as it can get overfitting to the dataset provided for training.

As for the case of SER, the dataset is often unbalanced, which is unavoidable since some specific emotions, such as contempt and fear, occur more rarely than neutral emotions in real life [1]. To address this issue, multi-task learning can be applied as it combines the training of the main tasks with other related tasks. In this way, the generalization can be improved by using the information learnt from other tasks [3]. Based on this knowledge, it may be useful to combine other classification tasks on voices, such as voice authentication, which intends to recognize the speaker's identity by analysing voice features. It is also a fact that expressions of emotion differ between individuals. Therefore, this project aims at exploring speaker-related characteristics of the sound.

This paper first goes through the existing work related to speaker identification and emotion recognition in section 2. After that, it shows the research questions and the approach taken to solve these questions in section 3. In section 4, the network models, the multi-task learning method and tools that are used for this project are demonstrated. Then follows section 5 where the dataset, implementation details and evaluation methods are explained. Then the results of the evaluation are presented in section 6. After that, section 7 gives a detailed discussion on the results. Finally, this paper draws a conclusion of the project and answers the research questions in section 8.

2. BACKGROUND

Currently, existing research on improving emotion detection with the identity of speakers [8] resulted positively as most of the testings showed an increase in accuracy of predicting the emotion of speakers. This research was conducted in two directions. The first one was to have additional speaker information combined in feature vectors as the input. The second conducted training separate emotion recognition models for each speaker and select the specific model according to the speaker when testing. This research also trained a network to identify the speaker and use this prediction in both directions. Although both models achieved the same highest increase of 10.2% on the accuracy of emotion recognition, using given identity information generally led to better improvement compared to training a separate speaker identifier.

There are also multi-task learning projects that combine SER with other auxiliary tasks such as detecting the gender, and naturalness of speakers [5]. The research examined this model under within-corpus and cross-corpus setting, where the former achieved significant gains for relatively larger corpora and the latter reported larger gains in most of the corpora.

Another research which used speaker identification to support emotion recognition was conducted for contextual conversation [6]. They aimed to improve the recognition of emotion by identifying speakers in a sequence of conversation, as human beings have consistency in their emotion to some extent. The results showed that training speaker identification as an auxiliary task improves the accuracy of emotion recognition in contextual speech.

However, the benefit of conducting multi-task learning with speaker identification in contextless speech, which is also an important field in SER, is still under-researched. To improve this research gap, this project works with a contextless dataset where the consistency of the emotion can not influence the testing results. The focus of the inputs is on the audio features extracted from the audio data, such as Mel-frequency cepstral coefficients (MFCC), instead of the context of the speech.

3. RESEARCH QUESTION

To achieve the goal mentioned in the previous chapter, this project aims to answer the following research question:

RQ1. In multi-task learning, to what extent do speaker identification task and speech emotion recognition task influence the accuracy of each other?

To conduct the research, we split this question into the

following sub-questions:
RQ1.1. How does training with speaker identification influence the accuracy of emotion recognition?
RQ1.2. How does training with emotion recognition influence the accuracy of speaker identification?
As figure 1 indicates, a set of experiments were conducted in order to answer the research questions. Data for both emotion recognition and speaker identification were prepared as subsets for training and testing. After that, the network model was specified. Then it was used in separate training for emotion recognition and speaker identification. Both tasks were evaluated on testing dataset. Another set of training and testing were conducted with deep multi-task learning, where the model was tasked with classification of emotions and speakers. Finally, the questions were answered based on the observed results.

4. METHODOLOGY

4.1 Network

The neural network used in this project is based on the VGGVox network presented by Albanie et al. [1]. The VGGVox network was developed by Nagrani et al. [7], who also established the Voxceleb dataset. As this project first attempts to use Voxceleb as the training dataset, the network and feature extraction are under the guidance of Albanie et al. [1]. The VGGVox is based on VGG-M [4] network, which has been proven effective for classification tasks on image data. To adjust this network for audio input, the fully connected layer fc6 of dimension 9×8 is replaced by a fully connected layer of 9×1 to support the frequency domain (the y-axis of input spectrograms) and an average pool layer of 1×11 which handles the length (x-axis) of the input spectrograms. Table 2 illustrates the full structure of the network.
The input of the network is a set of spectrograms extracted from the audio files. These audio files are first formulated into 4 seconds segments in order to keep the input effective and consistent in size. We apply zero-padding if the audio is shorter than 4 seconds, while if its length exceeds 4 seconds we use a random cut of it. Then after performing mean and variance normalisation, the Hamming window function is applied to provide a spectrogram of size 512×400 from the audio segment. This spectrogram is fed into the network as input and filtered by the layers presented in table 1. Because the research applies multi-task training for recognizing emotion and identifying speaker, the input is processed by different final layers for each task, respectivly fc8_1 for emotion recognition and fc8_2 for speaker identification, which are explained with more detail in section 4.2.

4.2 Multi-task training

As shown in figure 2, hard parameter sharing for MTL is applied in this project. The input is first processed by the shared layers of the network, which are the layers from conv1 to fc7 in table 1. Then the processed input goes through two separate fully connected layers, one for emotion classification and one for identity classification, providing two predicted labels. Since the dataset used for experiments contains 6 emotion classes and 91 speakers, the classification layers are linear layers with output dimension 6 for emotion recognition and 91 for speaker identification.
For both tasks, the model was trained using cross entropy

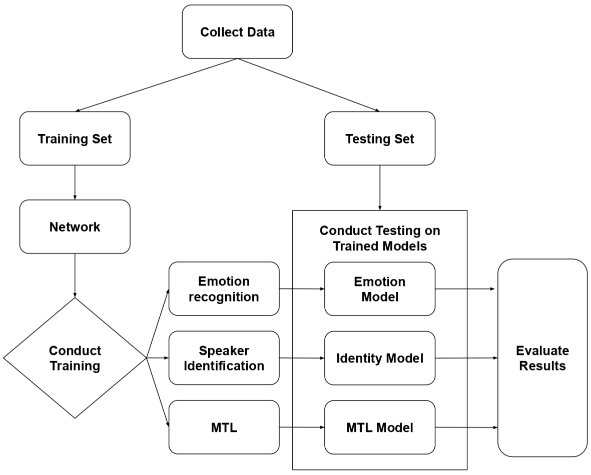

Figure 1. The approach taken for experiments.

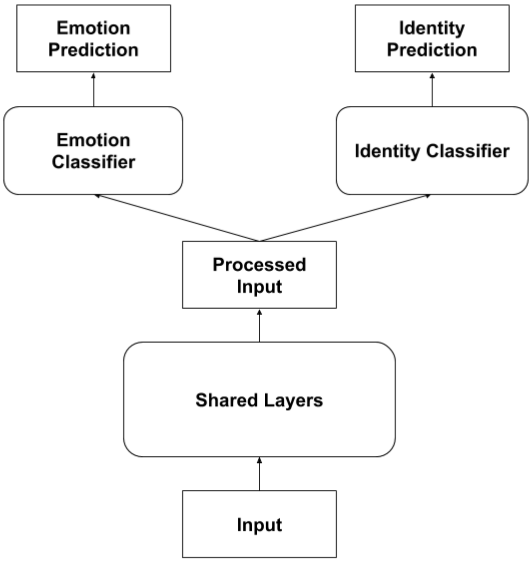

Figure 2. The structure of multi-task learning model.

Table 1. The architecture for the shared layers of the network.

Layer	Support	Filt dim.	# filt.	Stride	Data size
conv1	7×7	1	96	2×2	254×198
mpool1	3×3	-	-	2×2	126×99
conv2	5×5	96	256	2×2	62×49
mpool2	3×3	-	-	2×2	30×24
conv3	3×3	256	256	1×1	30×24
conv4	3×3	256	256	1×1	30×24
conv5	3×3	256	256	1×1	30×24
mpool5	5×3	-	-	3×2	9×11
fc6	9×1	256	4096	1×1	1×11
apool6	1×11	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8_1	1×1	1024	6	1×1	1×1
fc8_2	1×1	1024	91	1×1	1×1

loss function:

$$Loss(y, c) = -\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \qquad (1)$$

where M represents the number of classes, with $y_{o,c}$ denoting the binary indicator (0 or 1) if the observation o is labelled in class c and $p_{o,c}$ the predicted probability that observation o is of class c.
The calculation results for two tasks are then added with no weights specified to get the total loss for the entire MTL model, as shown below:

$$MTL\_loss = emotion\_loss + identity\_loss \qquad (2)$$

In this way, the total loss is be minimized instead of the separate loss for each task, providing an opportunity of improving generalization.

4.3 Programming Tools

This project uses Python as programming language and PyTorch as the library for machine learning. For features extraction, this project takes advantage of SciPy library to read the audio files and extract spectrograms as the input for training. The training is conducted on Colaboratory,[*] which provides an online platform and free access to GPUs for executing Python codes.

5. EXPERIMENT

5.1 Dataset

The experiment for this research is based on the CREMA-D dataset [2], which stands for Crowd Sourced Emotional Multimodal Actors Dataset. It contains 7442 original clips from 91 actors. The actors group containsreftable:dataset shows how the dataset is divided into training set and testing set.
Each clip in the dataset is labelled with one of the following emotions: anger, disgust, fear, happy, neutral, sad. The actors were asked to perform a few sentences with all six emotions. Therefore, the dataset has equally distributed labels, which provides a balanced training environment.
Figure 3 shows four sets of frame tracks and spectrograms extracted from clips of different actors expressing different emotions. As this research focuses on the audio features, only the spectrograms are concerned as the input vector during the experiments.

5.2 Implementation details

[*] https://colab.research.google.com

Table 2. Training and testing dataset of CREMA-D.

	Train	Test
Actors	67	24
Clips	5475	1967

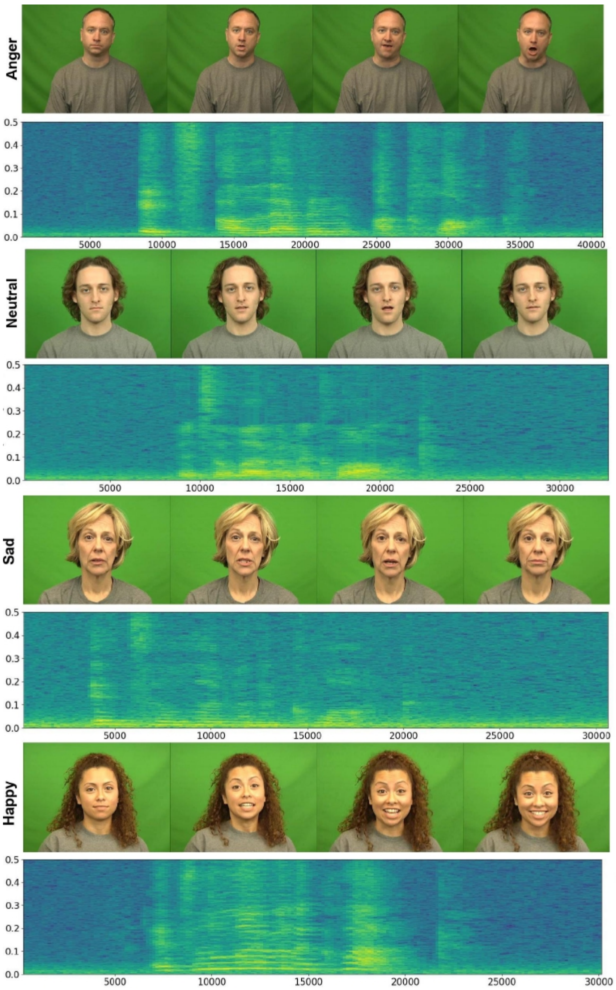

Figure 3. Sets of frame tracks and voice spectrograms where the x-axis of the spectrograms stands for time(sec) and the y-axis demonstrates frequency(Hz) of the audio segments.

Table 3. Accuracy on each emotion class (+ refers to the improvement when MTL applied)

	ANG	DIS	FEA	HAP	NEU	SAD	TOTAL
EMO	64.5%	37.1%	32.8%	43.1%	42.2%	47.3%	44.5%
MTL	72.5%	43.5%	42.0%	45.6%	51.8%	53.3%	51.5%
+	12.3%	16.2%	27.3%	7.0%	23.8%	12.8%	15.6%

For both single tasks and MTL, the models are trained for 60 epochs with batch size 64, using SGD optimizer with momentum of 0.9 and weight decay as 0.0005. The learning rate is set to 0.01 at the beginning of training, then multiplied by 0.1 after every 20 epochs. Although there are also experiments conducted with different parameters settings, the parameters described in this chapter achieve the best performance for emotion recognition. Therefore, the evaluation and discussion of the results in the following chapters are examined with this parameters setting.

5.3 Evaluation

The dataset is divided into 70% for training and the other 30% for testing. The testing set contains speakers with approximately same distributions on gender, age, race and ethnicity as the training set. For both emotion recognition task and speaker identification task, the models are evaluated by the accuracy of classifying the test data, which is calculated form the following equation:

$$accuracy = \frac{\#correct}{\#total} \tag{3}$$

where *#correct* denotes the number of observations that are classified into the correct classes, and *#total* represents the total number of observations that need to be classified.

6. RESULT

For emotion recognition task, both trained models for emotion task only and MTL are given the same testing set. The classifiers then predict emotion classes for each clip. Figure 4 provides the confusion matrices for both models. Furthermore, the accuracy for each emotion class as well as the improvement when applying the MTL model are given in table 3. As observed, both emotion and MTL models achieved the highest accuracy on Anger class and the lowest on Fear. The accuracies of detecting all of the emotion classes are increased when using the MTL model, with highest increase observed on Fear class, while the lowest on Happy. In total, applying MTL model has improved the accuracy for emotion recognition with 15.6%.
The speaker identity classifiers are examined by testing if the models could identify whether two speakers are the same or not. The result shows that the influence of applying the MTL model on the correctness for identification fluctuates according to the parameters for training. A few sets of experiments applying different parameters are conducted in order to achieve the best performance for emotion recognition task. When specifying the parameters as discussed in section 5.2, the accuracy on identification for single identifying model is 52.5% and 52.9% for MTL, showing an increase of 0.8%. However, when having the models trained on 50 epochs with an initial learning rate of 0.001 and decays to 0.0001 after 30 epochs, the influence of MTL is negative as the accuracy reduces from 53.1% to 52.8%.

7. DISCUSSIONS

Because of the differences in data and models used for training can strongly influence the evaluation, the comparisons in this chapter only focus on the works with similarity in either the dataset or the model.

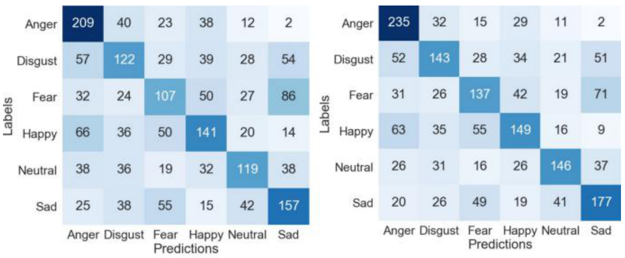

Figure 4. Confusion matrices for single emotion task model (right) and MTL model (left).

7.1 Performance of emotion recognition

In this project, the highest accuracy for training the emotion recognition task separately is 44.5%, outperforming the random classifier which has the accuracy of 16.7%. Comparing to the work of Albanie et al. [1] which provides the guideline on network structure, this project shortly falls behind its average accuracy of 44.7% (from 30.3% to 55.1% on different datasets). As for the dataset, CREMA-D is established by a project that involves human evaluators to rate the emotion from the audio. For audio-only data, which is used for this project as well, the results for ratings are 41% for matching, 46% for not matching and 13% for ambiguous. While the machine classifier proposed in this project has better performance on absolute emotion classification, these results shows the limitation of acted dataset since the actors and actresses may not express their emotions in a natural way as in real life. However, real life dataset also has its drawbacks, such as unbalanced data distribution. These features of the datasets make it difficult to evaluate a model directly. Another interesting observation was on the accuracy of each emotion classes. Both machine classifiers form this project and Albanie et al. [1] achieve best performance on recognizing Anger emotion, while the lowest recognition rates are both on Fear class. However, the human evaluators in CREMA-D project achieved best recognition rate on Happy class, on which the machine classifiers have relatively low accuracy. When it comes to recognizing Sad emotions, both machines significantly outperformed human evaluators.

7.2 Performance of MTL model

As mentioned in section 2, most of the current works which combine the emotion recognition with speaker information achieved positive results. In the work of Sidorov et al. [8], when applying separate models for each speaker, the improvement on accuracy reaches 10.2%, while in the work of Kim et al. [5] which applies MTL with the gender and naturalness of speakers has the highest improvement of 22.8%. The promotion of the model proposed in this project lies between them on 15.6%. Although the comparability of these results still needs discussion because of different parameters and datasets used in the experiments, it is observed that MTL models have advantages in combining speaker information. Moreover, the results imply that the number and correlation of the tasks in MTL models could be an important direction for improvement.

8. CONCLUSIONS

This paper proposes a model combining the emotion recognition task and speaker identification task using multi-task learning. The model is experimented on the CREMA-D dataset and achieves a relatively large gain of 15.6% on accuracy on emotion recognition, while slight fluctuation is observed on speaker identification. These results also provides answers to the research questions:

Answer to RQ1.1: Training with speaker identification provides significant promotion on the accuracy of emotion recognition.

Answer to RQ1.2: Training with emotion recognition does not make a significant difference in the accuracy of speaker identification.

For future development, the model proposed in this paper can be improved by specifying the weights for the loss of each task so that the loss function of MTL can better optimize the performance of emotion recognition. In this project, the training and testing datasets contain disjoint sets of speakers, which could have negative influence on speaker identification task. Therefore, to get a more direct observation for the performance of identifying speakers, another experiment could be conducted on training and testing dataset divided according to the presented sentences instead of speakers. Furthermore, testing the model with different network structures and datasets can also provide more reliable answers to the research questions. Combining the knowledge from the existing works, a direction to improve the project could be on increasing the number of related tasks that are more general. For example, changing speaker identification task into gender, age and language detection tasks. As mentioned in section 7.1, the performance of machine classifiers and human evaluators shows large differences between emotion classes. Therefore, future studies can conduct research on the difference between machine classification and human classification projects to explore a better way of feature extraction for speech emotion recognition.

9. REFERENCES

[1] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM ’18, page 292–301, New York, NY, USA, 2018. Association for Computing Machinery.

[2] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390, 10 2014.

[3] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 07 1997.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014*, 05 2014.

[5] J. Kim, G. Englebienne, K. Truong, and V. Evers. Towards speech emotion recognition ”in the wild” using aggregated corpora and deep multi-task learning. In *Interspeech 2017*, pages 1113–1117. International Speech Communication Association (ISCA), 2017.

[6] J. Li, M. Zhang, D.-H. Ji, and Y. Liu. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *ArXiv*, abs/2003.01478, 2020.

[7] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *ArXiv*, abs/1706.08612, 2017.

[8] M. Sidorov, S. Ultes, and A. Schmitt. Emotions are a personal thing: Towards speaker-adaptive emotion recognition. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4803–4807, 2014.